学 校 代 码 10459
学号或申请号 201012171872
密 级

郑州大学

硕士学位论文

虚词用法自动识别及其在依存句法分析中的应用研究

作 者 姓 名：张静杰
导 师 姓 名：昝红英 教授
学 科 门 类：工 学
专 业 名 称：计算机软件与理论
培 养 院 系：信息工程学院
完 成 时 间：2013 年 5 月

A thesis submitted to

Zhengzhou University

for the degree of Master

Studies on Automatic Recognition of Functional Words Usages and Application on Dependency Parsing

By Jingjie Zhang

Supervisor: Prof. Hongying Zan

Computer Software and Theory

College of Information and Engineering

May 2013

摘　要

现代汉语中词语主要分为实词和虚词两大类，其中虚词包含副词、介词、连词、语气词、方位词、助词等。虚词不能充当句法成分，但用法比较复杂多样。同一个虚词在不同的上下文中词性不同，即使词性相同，在不同语境中的用法也可能不同。因此需要对虚词的各种用法进行具体的分析和研究，为文本的理解提供方便，也有利于现代汉语的深入研究。

本文在构建的“三位一体”虚词用法知识库的基础上，以副词为例，分别采用了基于规则和基于统计的方法对现代汉语副词用法进行自动识别的研究，其中在基于统计的方法中分别使用了条件随机场模型、最大熵模型和支持向量机模型进行研究分析。实验结果表明，基于统计的副词用法自动识别的效果在总体上要优于基于规则的方法，而且三种统计模型中以支持向量机模型的效果最好，但从单个用法的识别角度分析，一些用法在基于规则的方法上识别的效果较好。因此，本文结合基于规则和基于统计两种方法的优点，提出了规则和统计相结合的思想。实验结果表明，采用规则和统计相结合的方法在副词用法自动识别研究中取得较好的效果。

在虚词用法自动识别研究的基础上，本文分析了其在汉语依存句法分析中的应用。对汉语依存句法进行分析时，采用哈尔滨工业大学的 HIT-IR-CDT 树库以及语言技术平台 LTP，共有 24 种依存关系。通过对 LTP 中依存句法分析的功能进行详细分析，发现其中的并列关系识别效果较差。本文对并列关系中的标注情况进行了分类总结，根据连词用法识别出句子中的并列结构短语，根据识别结果对依存句法分析的结果进行处理，从而提高汉语依存句法分析中并列关系的识别效果。实验结果表明，采用并列结构信息后，并列关系的识别效果明显提高。

关键词：虚词用法　自动识别　规则与统计结合　并列关系　依存句法分析

Abstract

Modern Chinese words are mainly divided into two categories of content words and functional words, and functional words include adverbs, prepositions, conjunctions, modal words, location words, and auxiliary. Although the functional words can not act as syntactic components, its usages are more complex and diverse. The same functional word may have different parts of speech in different contexts, even if they have the same part of speech, their meanings and usages may different. Therefore it is necessary to analysis and research on various usages of functional words, and it not only offers facilities for the understanding of the text, but also makes for the in-depth study of the modern Chinese.

In this paper, it builds the "Trinity" knowledge base of modern Chinese functional words. On this basis the modern Chinese functional words usages automatic recognition, take adverbs for instance, are researched with the rule-based and the statistics-based approach, which uses conditional random field model, maximum entropy model and support vector machine model for research and analysis. The experimental results show that the statistics-based approach is in general better than the rule-based method, and the support vector machine model has the best result in three statistical models. However, from the analysis of single usage recognition, the rule-based approach is better on some usage. Therefore, according to the advantages of both rule-based and statistics-based approach, this paper proposes the idea of combining rules and statistics. The experimental results show that this method achieved good results in the automatic recognition research on adverbs usages.

On the basis of functional words usages automatic recognition, this paper analyzes its applications on Chinese dependency parsing. On Chinese dependency syntax analysis, it uses HIT-IR-CDT Treebank and language technology platform LTP of Harbin Institute of Technology, which contains 24 dependencies. It founds that the recognition effect of coordination relations is poor after the detail analysis about the LTP. This paper sums up the labels of coordination relations, and then recognizes the

parallel structure phrases of the sentence with the conjunction usages. Finally, the recognition results of the dependency parser can be processed with the parallel structure phrases to improve Chinese dependency parsing. The experimental results show that the recognition effect of coordination relations is obviously improved after using the parallel structure information.

Keywords: functional words usage; automatic recognition; rules and statistics; coordination relations; dependency parsing

目　录

图表目录

图目录

表目录

1 引言

1.1 研究意义

汉语中的词语分为实词和虚词两大类别，实词主要包括形容词、名词、动词等，虚词一般由连词、副词、介词、语气词、方位词、助词等组成。从功能的角度分析，实词含有实际的意义，可以在句子中充当主语、谓语、宾语等句子成分，而虚词不能充当任何句法成分。从意义上来看，实词表示时间、事物、动作、处所等内容，虚词只有语法意义，表示某些逻辑概念或者起到一些语法作用。

虚词在汉语中有着非常重要的地位，中文的句法手段主要是词序和虚词[1]。汉语既没有蒙语、日语等语言里的黏附形式，也没有英语、俄语、法语等语言中的形态标志和屈折变化[2]。因此，在汉语中无法通过语言的黏附形式、形态标志、屈折变化等信息表达的语义和语法任务，常常由虚词来完成。从这个角度分析，对现代汉语中的虚词进行研究，对中文的语言处理和语言理解有重要作用。

本文中广义虚词主要包含连词、副词、介词、语气词、方位词、助词等。同一虚词在不同的上下文中可以显示不同的词性，在句子中的语法意义不同，所表达的意思自然也不同，如以下例句：

（1）他们组建了自己的运输公司、营销队伍，和市场直接接轨。

（2）贵州南部、江南、华南西部和华北将有小到中雨。

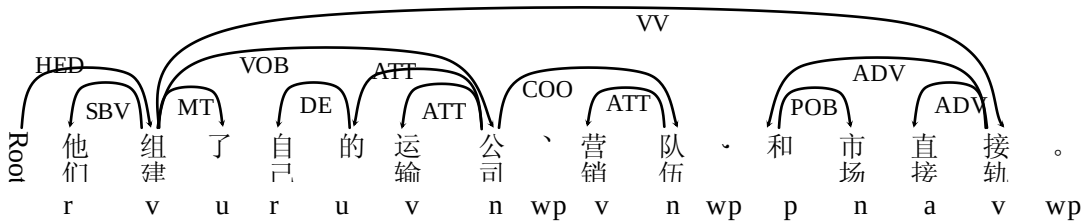


图 1.1 例句（1）的依存句法分析结果

例句（1）的依存句法分析结果如图 1.1 所示，句子中的“和”为介词，表示与某种事物有联系，介词“和”与名词“市场”构成介宾关系 POB，之后与动词“接轨”构成状中结构 ADV；例句（2）的依存句法分析结果如图 1.2 所示，句

子中的“和”为连词，用于连接结构或者类别相同（或者相近）的并列成分，表示平等的联合关系，将句子中的“贵州南部”、“江南”、“华南西部”、“华北”等成分联系起来组成并列关系，充当句子的主语。

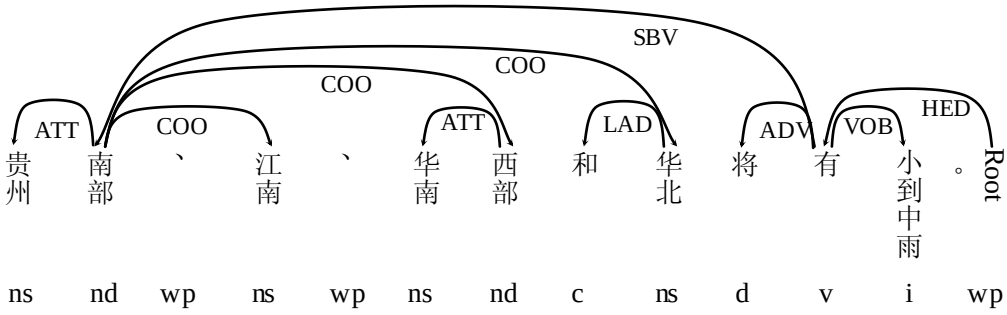


图 1. 2 例句（2）的依存句法分析结果

同一虚词在词性上的不同可以通过分词和词性标注进行区分，而且目前关于分词和词性标注的研究已经相对成熟，可以解决一些因词性不同而导致的语义理解上的问题。在语料中将上述例句中的“和”分别标注为“和/p”和“和/c”，再次对句子进行理解分析时，便可以对“和”的两个词性进行区分。

同一虚词的同一词性可能因为上下文语境的不同而表达不同的意思，即用法不同，如下述例句：

（3）**在**草地的中间有一个水池。

（4）我**在**学习上很努力。

（5）我**在**他的帮助下完成了工作。

这三个例句中的“在”均为介词，但用法各不相同。例句（3）中介词“在”用在形容词、动词或者主语前，表示事物存在或者动作发生的处所；例句（4）中的介词“在”也用在形容词、动词或者主语前，表示范围；例句（5）中的介词“在”用在动词或者主语前，表示条件。

由介词“在”构成的介词结构在三个例句中作状语等成分修饰句子的核心词，例句（3）的介词结构为“在草地的中间”（图 1.3），表示处所；例句（4）的介词结构为“在学习上”（图 1.4），表范围；例句（5）的介词结构为“在他的帮助下”（图 1.5），表示条件。如果可以识别出这些介词结构，对句子的理解有很大的帮助。

即使将三个例句中的介词“在”都加上词性标注信息，即“在/p”，也不能表达出例句意思和意义上的不同，因为介词结构的不同是因介词用法的不同而

体现出来的。由此，仅仅依靠分词和词性标注信息并不能解决此类问题。

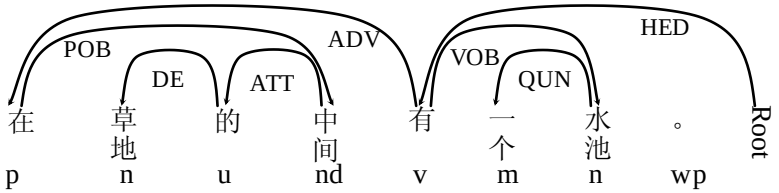


图 1.3 例句（3）的依存句法分析结果

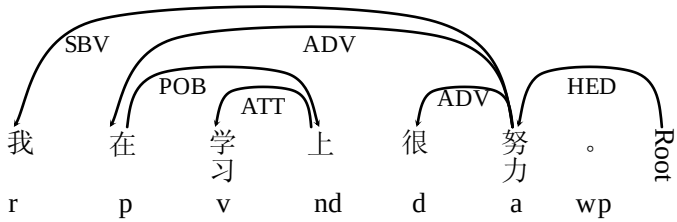


图 1.4 例句（4）的依存句法分析结果

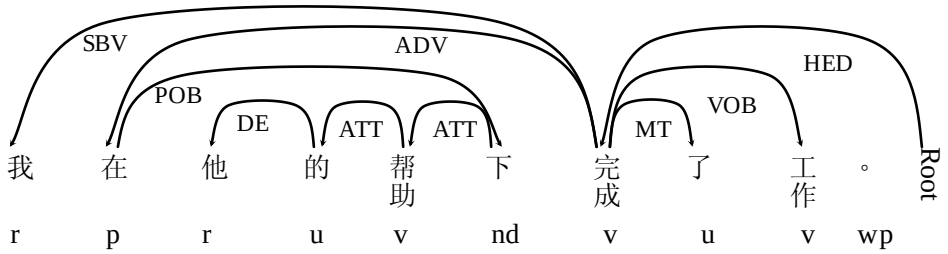


图 1.5 例句（5）的依存句法分析结果

为了解决上述问题，需要将虚词在词性标注的基础上，再进行细节上的划分，即为虚词标注上用法信息。这需要对语料进行大规模的分析，对其中的虚词进行细致全面的研究，总结出一定的规则，结合人工和机器学习两者共同的作用对虚词用法进行自动识别的研究，并分析虚词用法在句法分析中的应用。

1.2 研究的背景

郑州大学自然语言处理实验室于 2010 年承担了国家自然科学基金项目（60970083）“规则与统计相结合的现代汉语虚词用法自动识别研究”，在此之前，郑州大学已经完成了北京大学计算语言学教育部重点实验室开放课题基金资助项目（KLCL-1004）“现代汉语虚词知识库研究及大规模虚词用法标注语料库的构建”，国家 973 项目“文本内容理解的数据基础”（2004CB318102）的子任务“现代汉语虚词用法标注语料库的校对与扩充”，以及河南省科技创新人才

杰出青年基金项目（104100510026）“面向文本内容理解的现代汉语虚词知识库研究”。本文中关于虚词用法自动识别的研究就是上述工作的一部分，关于虚词用法在依存句法分析中的应用研究是中科院自动化所模式识别国家重点实验室开放课题基金项目“虚词用法在文本理解中的应用研究”的相关内容。

汉语中的词分为实词和虚词两大类，两者在对汉语句子篇章的理解上都有非常重要的作用。而目前国内外的研究学者对汉语的研究主要针对实词，针对虚词的研究大多停留在面向人用的词典之类的非计算机领域，涉及到虚词语言处理的技术和研究很少。

针对这种情况，俞士汶、朱学锋等[3]提出“三位一体”建设现代汉语广义虚词知识库的想法和思路，其中的广义虚词包括连词、语气词、副词、助词、介词和方位词等等。昝红英等[4]对现代汉语虚词的用法进行了大量的分析和研究。在这些思想和研究的基础上，刘锐[5]采用基于规则的方法对副词用法进行了自动识别的研究，张军珲[6]等采用了基于统计的方法对常用汉语副词进行了用法自动识别研究，袁应成等[7]对现代汉语的介词短语边界识别进行了研究，韩英杰等[8]采用了基于规则的方法对现代汉语的常用助词进行了用法自动识别研究，周溢辉[9]针对现代汉语虚词中的语气词进行了用法自动识别研究，周丽娟[10]研究了连词用法以及连词结构短语的自动识别。

本文在以上研究的基础之上，完善了现代汉语虚词中副词的用法规则库、用法词典以及用法语料库，以基于规则与基于统计方法的副词用法自动识别[11]为基础，分析两种方法的优点和不足，将两者的优点进行结合，提出采用规则和统计相结合的方法对副词的用法进行自动识别。最后，在虚词用法自动识别研究的基础上，分析了虚词用法在依存句法分析中的应用，其中主要针对依存关系中与虚词关系比较密切的并列关系进行分析。

1.3 本文研究工作

本文在已经构建的现代汉语虚词知识库的基础上，对虚词用法进行了自动识别研究，并对虚词用法知识库的应用进行了分析探讨，即将其应用到依存句法分析中。

本文的主要研究工作包括：

（1）根据现代汉语虚词用法规则以及现代汉语虚词用法词典对虚词进行基于

规则的用法自动识别研究。首先采用袁应成已经开发的规则标注系统[12]对北京大学计算语言学研究所提供的《人民日报》2000 年 1 月到 6 月分词和词性标注后的语料[13]进行用法标注，然后对标注的结果进行人工校对，完善虚词用法语料库，之后对比校对前后的语料，分析规则标注错误的部分，对规则进行优化，使其更能全面的表达虚词的用法，完善虚词用法的规则库。

（2）根据完善后的虚词用法语料库对虚词进行基于统计的用法自动识别研究。本文采用了三种统计模型进行实验，主要对在序列标注方面占优势的条件随机场模型进行详细的分析，并将其与基于规则的方法进行对比，分析各自的优点和不足。

（3）在基于规则和基于统计两种方法的分析研究后，结合各自的优点提出了规则和统计相结合的方法，并对其进行了实验研究分析，改善虚词用法的识别效果。

（4）对虚词用法的应用进行探讨。将由虚词用法识别出的连词结构短语应用到依存句法分析中，以提高并列关系的识别情况。首先对哈尔滨工业大学提供的语言技术平台 LTP 中的依存句法分析模块进行了详细的分析，总结出并列关系的标注情况；然后将语料中表并列关系的连词进行用法标注和并列结构短语标注；最后利用并列结构标注信息对句法分析的结果进行优化，提高依存句法分析中并列关系的识别效果。

1.4 论文结构框架

根据本文的主要研究内容，本文主要将其分为五章进行阐述。各个章节的具体安排如下：

第 1 章，引言。简单介绍了本文的研究背景和意义，主要的研究工作以及论文的结构框架。

第 2 章，相关研究。主要介绍了“三位一体”的现代汉语虚词用法知识库的构建，包括现代汉语虚词用法词典、现代汉语虚词用法规则库以及虚词用法语料库。介绍依存句法分析的研究方法以及汉语依存句法分析的研究现状。

第 3 章，现代汉语虚词用法自动识别。以虚词中的副词为例，首先介绍了基于规则的副词用法自动识别，并详细分析了规则的优化过程以及实验情况；然后介绍了基于统计的副词用法自动识别，分别采用了条件随机场 CRF[14]、最大

熵 ME[15]、支持向量机 SVM[16]三种统计模型，并对每种模型进行了实验分析；最后根据对基于规则和基于统计方法的分析，提出了规则和统计相结合的思想，并进行了实验分析。

第 4 章，虚词用法识别在依存句法分析中的应用。将由虚词用法识别出的连词结构短语应用到依存句法分析中，以提高并列关系的识别情况。

第 5 章，结论和展望。对本文的研究工作进行总结，并提出下一步的研究思路和方向。

2 相关研究

2.1 现代汉语虚词用法知识库

虚词在句子中不能单独充当句法成分，有连接或者依附于各类实词的语法意义。根据虚词在句子中的语法意义以及与实词的搭配情况，可以将虚词分为语气词、副词、连词、介词、助词等等。汉语是虚词特别丰富的语言，这主要由汉语本身的特性所决定。

现代汉语中虚词的数量虽然很少，但是总体上来说，虚词的重要性并不亚于实词，对于一些意义纷杂的虚词来说，其重要性甚至超过实词。因为虚词在应用中往往影响着整个句子的结构，进而影响我们对整个句子甚至整个段落的理解。如果虚词运用的精准，不仅可以提高我们对虚词所在句子的理解效果，而且常常会帮助我们理解整个段落甚至整篇文字。因此，对现代汉语虚词进行研究非常必要。

同一虚词可能显示不同的词性，同一虚词的同一词性也可以有不用的用法，表达不同的意思。在对虚词研究的很长一段时间里，研究的成果大部分都是面向人用的，对虚词的个性描写很难直接应用到自然语言处理的研究中，在一定程度上影响了机器对文本的理解，限制了自然语言处理领域的研究和发展。

为了解决上述问题，实现面向机器的虚词研究问题，昝红英等从计算语言学的理论观点出发，在对真实的语料进行虚词用法规律考察的基础上，构建了面向机器的现代汉语虚词用法信息词典和虚词用法规则库[17]，并对《人民日报》分词和词性标注语料中的虚词用法进行了研究考察，构建了面向自然语言处理的现代汉语广义虚词语料库，实现了俞士汶等提出的“三位一体”构建现代汉语广义虚词知识库的思想[18]。

2.1.1 现代汉语虚词用法词典

昝红英等构建的现代汉语虚词用法知识库中包含所有的广义虚词，即介词、连词、方位词、副词、助词、语气词等。其中方位词可以说是名词的一个附类，主要跟在介词的后面共同组成介词短语结构，也可以附着在名词后表示处所或时间；副词能够修饰动词或者形容词，但是不能修饰名词；介词不能在句子中

单独充当谓语，常常与代词、动词、名词等搭配组成介词短语结构；连词主要将语义相关的语言单位连接在一起，反应事物之间的关系；语气词主要用于句子末尾，表达各种语气。虚词的用法和作用各不相同，这就使得必须对每个虚词的每种用法进行详细的分析和研究。

表 2.1 副词“都”部分属性用法词典样例

ID	释义	用法
d_dou1_1	表示总括全部。<b>	除问话以外，所总括的对象必须放在“都”前。也可以说“全都”，总括的意思更明显。<b>
d_dou1_1a	表示总括全部。<b>	所总括的对象可以用表示任指的疑问指代词。<b>
d_dou1_1b	表示总括全部。<b>	所总括的对象前可以用连词“不论、不管、无论、凡是、只要”。 <b><z>
d_dou1_1c	表示总括全部。<b>	问话时总括的对象（疑问代词）放在“都”后。<b>
d_dou1_1d	表示总括全部。与“是”搭配。说明原因，有责备的意思。<b>	与“是”搭配。<b>
d_dou1_2	甚至。“都”轻读。<b>	修饰动词或动词短语。<z>
d_dou1_2a	甚至。“都”轻读。与“连”字同用，有强调语气的作用。<b>	与“连”字同用。<b>
d_dou1_2b	甚至。“都”轻读。<b>	“都”前后用同一个动词（前肯定，后否定）。<b>A~(不|没|没有|未|df)A 或 A~A(不|没|没有|未|df)。<z>
d_dou1_2c	甚至。“都”轻读。<b>	一+量词+...~+动词（否定式）。<b>
d_dou1_2d	甚至。“都”轻读。<b>	用于表示让步的小句，引出表示主要意思的小句。<b>
d_dou1_3	已经。<b>	句末常用“了”。 <b>

根据张斌的《现代汉语虚词词典》（<x>）、吕叔湘的《现代汉语八百词》（<b>）、《现代汉语词典》（第五版）（<h>）、1998 年 1 月《人民日报》的分词和词性标注语料库的统计分布（<r>）以及其他的语法学家相关论著（<z>）（“<>”是用法信息词典中的来源标记）对现代汉语中的每个虚词的义项及用法进行了详细的分析和总结，最终构建了现代汉语虚词用法信息词典，以副词“都”为例，其在虚词用法信息词典中的描述如表 2.1 所示，副词“都”有 3 个义项 11 个用法。表 2.1 中的“ID”是用法编码属性，编码规律为：词性_汉语全拼和声调_义项编码和用法编码。例如：d_dou1_1c，表示副词（d）“都”（dou）为平声（1）

的 1 义项的 c 用法。“释义”表示语义的信息，相同的释义可以有不同的用法描述。“用法”对释义进行细节的划分，对词语表达的意义理解地更清晰。

2.1.2 现代汉语虚词用法规则库

由表 2.1 可以看出，现代汉语虚词用法词典中的用法描述等信息也属于面向人用的用法描述，很难直接应用到自然语言处理的研究过程中，因此我们需要将人类能够识别的现代汉语虚词用法词典中的用法描述转化为机器可以识别的用法描述。咎红英等对现代汉语虚词用法词典中的用法描述进行了 BNF 范式[19]的形式化描述。

形式化描述过程中抽取的现代汉语虚词用法特征主要包括：句首 F，表示出现在句首的词性或者词语信息；左搭配 M，表示与目标虚词左侧搭配出现的词语或者词性信息；左紧邻 L，表示与目标虚词左侧紧邻出现的词语或者词性信息；右紧邻 R，表示与目标虚词右侧紧邻出现的词语或者词性信息；右搭配 N，表示与目标虚词右侧搭配出现的词语或者词性信息；句末 E，表示目标虚词所在的句子中句末出现的词语或者词性信息。除了这些主要的特征描述，在规则库中还要引入一些其他的符号[12]，尽量使得规则库中的符号信息能够更加全面的表述虚词词典中的用法描述信息。

现代汉语虚词用法规则的描述形式：

<ID>→[F][M][L][R][N][E]

F→<词语 1>|<词语 2>|…|n|v|a|…

M→<词语 1>|<词语 2>|…|n|v|a|…

L→<词语 1>|<词语 2>|…|n|v|a|…

R→<词语 1>|<词语 2>|…|n|v|a|…

N→<词语 1>|<词语 2>|…|n|v|a|…

E→<词语 1>|<词语 2>|…|n|v|a|…

其中规则元语言“→”意思是“定义为”，中括号“[ ]”中表示的是可选内容，“|”表或运算。根据这种规则描述方法，表 2.1 中副词“都”的用法描述对应的用法规则可以表述为：

$都

@<d_dou1_1>→N ^N→v|a

@<d_dou1_1a>→M ^M→谁|哪里|什么|怎么|哪儿|哪|<ry>|<ryw>

@<d_dou1_1b>→M ^M→(不论|不管|无论|虽然|尽管|凡是|只要)*{，}

@<d_dou1_1c>→NE ^N→谁|哪里|什么|怎么|哪儿|哪|<ry>|<ryw> ^E→？

@<d_dou1_1d>→FR ^F→~ ^R→是

@<d_dou1_2>→N ^N→v

@<d_dou1_2a>→M ^M→连|甚至

@<d_dou1_2b>→A~A(不|没|没有|未|<df>) ^A→v

@<d_dou1_2b>→A~(不|没|没有|未|<df>)*A ^A→v

@<d_dou1_2c>→MN ^M→一q ^N→不|没|没有|未|<df>

@<d_dou1_2d>→N ^N→[，]*(不|没|<df>)

@<d_dou1_3>→E ^E→了，

2.1.3 现代汉语虚词用法语料库

在现代汉语虚词用法词典和用法规则库的基础上，对初始语料进行用法标注，形成现代汉语虚词用法语料库，为虚词用法的研究提供平台和依据。这里的初始语料是北京大学计算语言学研究所提供的 2000 年 1 月至 6 月《人民日报》分词和词性标注语料[13]，用法标注前后的语料对比如下：

用法标注前的语料：

保留/v 军队/n ，/wd 领导人/n 可以/vu 参加/v 国家/n 的/ud 管理/vn ；/wf **在/p** 一个/mq 中国/ns 的/ud 原则/n 下/f ，/wd 什么/ry 问题/n **都/d** 可以/vu 谈/v 等/u ，/wd 符合/v 台湾/ns 人民/n 求/v 稳定/a 、/wu 求/v 和平/a 、/wu 求/v 发展/v 的/ud 期望/vn ，/wd **也/d** 符合/v 台湾/ns 人民/n 的/ud 心愿/n 。/wj

用法标注后的语料：

保留/v 军队/n ，/wd 领导人/n 可以/vu 参加/v 国家/n 的/ud 管理/vn ；/wf **在/p<p_zai4_4>** 一个/mq 中国/ns 的/ud 原则/n 下/f ，/wd 什么/ry 问题/n **都/d<d_dou1_1a>** 可以/vu 谈/v 等/u ，/wd 符合/v 台湾/ns 人民/n 求/v 稳定/a 、/wu 求/v 和平/a 、/wu 求/v 发展/v 的/ud 期望/vn ，/wd **也/d<d_ye3_1b>** 符合/v 台湾/ns 人民/n 的/ud 心愿/n 。/wj

对语料中包含的所有虚词均标注用法信息后，便可得到现代汉语虚词语料库，与现代汉语虚词用法词典、现代汉语虚词用法规则库共同构成现代汉语虚

词知识库，为虚词的用法识别以及应用提供可靠的依据和平台，最终促进信息抽取、句法分析等自然语言处理相关领域的发展。

2.2 依存句法分析

自然语言处理的分析技术可以分为两个层面：浅层分析和深层处理。其中浅层分析一般只对句子的局部内容进行分析和处理，例如分词、词性标注等，目前这些技术已经取得了很好的成绩，并应用到文本分类、信息抽取等其他自然语言处理领域，而关于句法分析、语义分析等对语言深层处理的研究技术还不成熟，还有待更深入的研究。

句法分析是自然语言处理中的关键技术之一，其基本任务是确定句子的句法结构，一般而言，获取句子中的句法结构不是自然语言处理的最终目标，但它往往是实现具体目标的重要环节，甚至在某些情况下是非常关键的一步，目前汉语句法分析与其他西方语言的分析效果相比还有一定的差距，主要是因为汉语具有许多自身的特点，因此，研究和开发适合汉语的句法分析技术便是众多专家关注的问题[20] [21] [22]。

在自然语言处理领域，句法分析研究中遵循的语法体系主要分为两个方向，短语结构语法和依存语法。

（1）钱其琛为中国驻南非大使馆揭开馆牌。

以例句（1）为例，基于短语结构语法的句法分析结果如图 2.1a 所示，句法分析树中包含非终结点、短语标记和终结点三部分组成。这种句法树的形式直观表达了每个词语在句子中的成分以及与其它词语之间的内在联系，帮助我们理解整个句子的结构，而有时我们不需要或者不仅仅要知道整个句子的短语结构树，而且要知道句子中词语与词语之间的依存关系，进而方便我们理解整个句子的意思，这就需要采用基于依存语法的句法分析。对同一例句进行依存句法分析后的结果如图 2.1b 所示。

与短语结构句法相比较，依存句法有以下的优点：形式简单，只包含句子中的词汇、依存弧以及依存关系，而没有增加额外的非终结符号，这也使得我们更方便更容易理解整个句子的意思；侧重反映语义关系，可以很容易和语义分析结合；也可以用于表示交叉关系，因此适用于大多数语言；有利于实现线性时间的搜索算法。因此，依存句法分析受到国内外众多自然语言处理研究者的

广泛关注[23]。

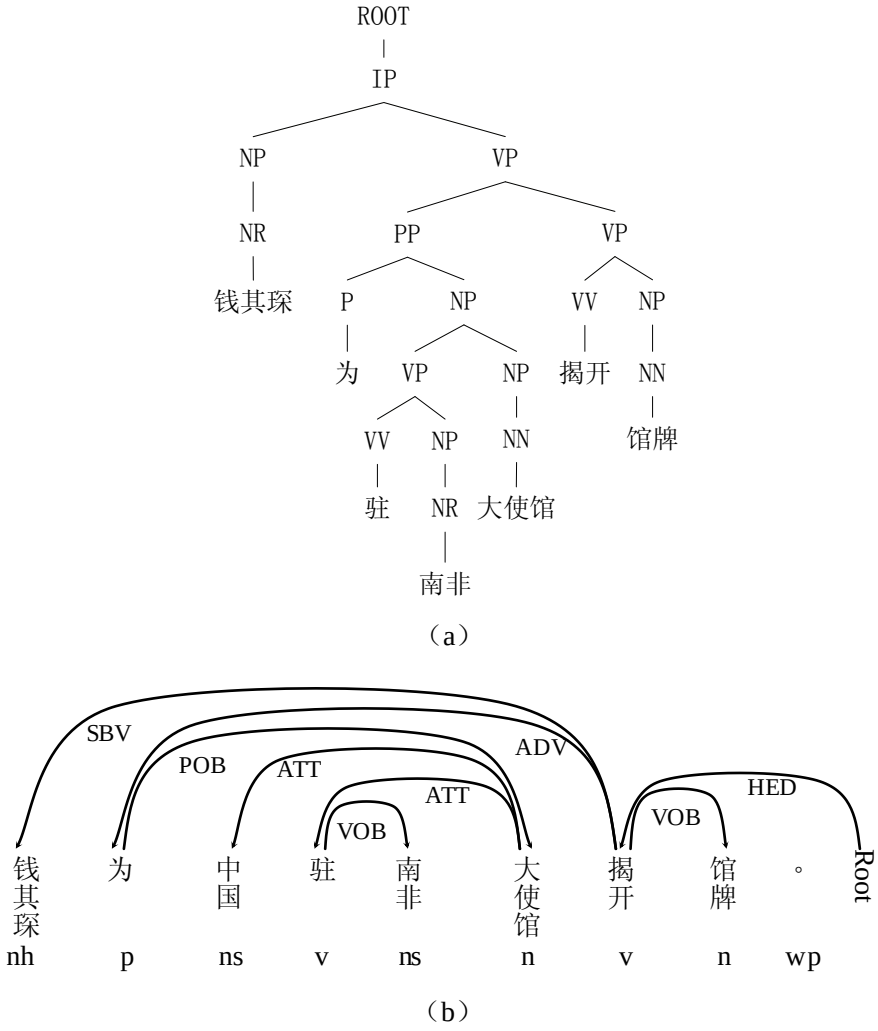


图 2.1 例句（1）的句法分析结果

2.2.1 依存句法分析的研究

依存语法理论是由法国语言学家 Lucien Tesniere 创立的，该理论认为，一切结构句法现象可以概括为关联、组合和转位三大核心，句法关联建立的是词与词之间的从属关系，这种从属关系是由从属词和支配词联结而成；动词是句子的中心并支配别的成分，本身并不受其他成分的支配[24]。

之后计算语言学家 J.Robinson 在论文“依存结构和转换规则”中提出了依存语法的四条公理[25]：

（1）一个句子只有一个独立的成分；

（2）句子的其他成分都从属某一成分；

（3）任何一个成分都不能依存于两个或两个以上的成分；

（4）如果成分 A 直接从属于成分 B，而成分 C 在句子中位于 A 和 B 之间，那么，成分 C 或者从属于 A，或者从属于 B，或者从属于 A 和 B 之间的某一成分。

依存句法理论的提出引起了众多自然语言处理研究者的注意，并将其应用到各自的研究领域。在我国，冯志伟首先将依存句法理论引入自然语言处理研究[26]。

句子的依存句法结构图用带有方向的弧来表示两个成分之间的依存关系（图 2.1b），支配者在有向弧的发出端，被支配者在箭头端，即被支配者依存于支配者。依存弧上显示的是两个词语之间的依存关系。

自依存句法理论提出以来，关于依存句法分析算法的研究有很多，目前主要归为四类[20]：生成式的分析方法、判别式的分析方法、决策式的分析方法和基于约束满足的分析方法。

生成式的分析方法主要是采用联合概率模型生成一个句子的多个依存句法树，并赋予每个依存句法树一个相应的概率值，针对不同的依存句法树，计算其相应算法的概率总值，概率总值最大的便为这个句子最终的依存句法树。Eisner 给出了第一个生成式的依存句法分析模型[27]，该模型是由短语结构句法分析算法直接演变而来的。Collins 等人也对生成式分析方法进行了研究[28]。生成式分析方法与短语结构句法分析的关系非常密切，在依存语法理论提出后的很长一段时间里，关于依存句法的分析主要是由短语结构方法演化而来的。

判别式句法分析主要采用的是条件概率模型，避开联合概率模型中要求的独立性假设。McDonald 等人[29]实现的最大跨度树是典型的判别式分析方法。最大跨度树（Maximum Spanning Tree, Mst）模型将求解一个句子的最佳依存句法分析树转化为寻找该句子最高打分的依存句法分析树的问题。这种方法中应用了很多的运筹学方法以及机器学习方法，在可计算性上占有很大的优势。在该方法的研究过程中，其主要任务放在降低算法的复杂度上面。

决策式句法分析是以指定的方向逐步读取一个待分析的词语，并为该词语产生一个单一的分析结果，直至该词语序列的结尾。与其他方法在每一步的分析过程中产生多个候选结果不同，该方法在句法分析的每一步只保留一个当下最优的分析结果。因此该方法分析得到的语法树并不是全局最优的结果，这使得

在降低算法复杂度的同时，也降低了句法分析结果的准确率。所以在保证算法确定性的前提下提高分析结果的准确率便成为该方法的研究重点。

基于约束满足的句法分析方法将依存句法分析看作是可用约束满足问题来表达的有限构造问题。根据已规定好的约束对依存分析树进行修剪，将不合约束的部分剪掉，直至最终满足约束的依存分析树。

这四种方法各有优势和不足，本文使用的哈工大社会计算与信息检索研究中心研制的语言技术平台（Language Technology Platform, LTP）[30]中的依存句法分析是根据汉语的特点，对判别式分析方法中的 Mst 模型进行实现。

在句法分析的研究中，英语的研究开始的最早，研究的效果也最好。在早期基于短语结构语法的句法分析一直占据主流，依存语法理论及其研究和实践相对比较晚。Eisner 最先将宾州树库（Penn Treebank）转化为依存的表示形式，然后进行基于依存语法的句法分析研究[27]。

随着依存句法分析研究的深入，引起了越来越多的自然语言研究者的广泛关注，与此同时也出现了很多关于依存句法分析的算法研究。McDonald 等人[31] [32]的研究得到了很多研究者的认同，他们不但提出了 Mst 模型进行依存句法分析，并对该方法进行不断的改进，提高算法的准确率。目前以宾州树库作为语料信息，基于依存语法的分析结果与基于短语结构的分析结果已经接近[31] [32]，这在一定程度上对基于依存语法的句法分析方法也是一种肯定。

2.2.2 汉语依存句法分析的研究现状

随着依存句法分析在英语等其他语言的研究上取得优异的成绩，近些年，国内外的学者对汉语依存句法分析的研究也产生了浓厚的兴趣。但是依存句法分析在汉语中的研究比较晚，与英语的研究相比效果也比较差。这主要是由汉语本身的特性所造成的，汉语中的字、词语以及短语之间没有明显的界限标记，汉语在研究过程中也无法形成自己的语法体系。

Zhou[33]是最早对汉语依存句法分析进行研究的学者之一，他提出对汉语的依存句法分析要采用分块的思想，即应用一些制定好的语法规则，先对待分析的句子进行分块处理，找出句子中关系固定的语块，之后对整个句子进行依存句法分析。王蕾[34]采用决策式的依存句法分析算法对汉语依存句法分析中的长句依存进行了研究，并针对 Arc-eager 决策式依存句法分析算法中所出现的 Early-reduce 问题进行了改进，以提高算法的性能。袁里驰[35] [36]利用了语义、语

法等知识建立了一种基于依存关系的句法分析统计模型，之后又利用语义、语法等知识对中心词驱动的句法分析模型规则进行分解和修改，结合分词和词性标注进行句法分析，提出一种可同时考虑多个语义依存关系的模型。刘海涛[37]等利用了语言学的手段对汉语依存句法进行了分析，提高了依存句法分析的结果。对汉语依存句法分析的研究涉及了依存句法分析的各种方法，同时也加入了汉语本身的特性，如汉语的语法和语义等知识，这些研究都取得了很好的效果，为之后句法分析的深入研究以及应用发展提供了很好的前提条件。

在汉语依存句法分析研究中，以哈工大社会计算与信息检索研究中心的研究最为全面典型，其不仅研究了依存句法，还建立了依存树库，为之后的研究提供了平台。

2.3 本章小结

本章主要介绍了现代汉语虚词用法知识库和依存句法分析的相关内容。其中现代汉语虚词用法词典、现代汉语虚词用法规则库和现代汉语虚词用法语料库，这三部分构成了“三位一体”现代汉语虚词用法知识库。在依存句法分析部分主要介绍了依存语法理论，依存句法分析的研究方法，以及汉语依存句法分析的相关研究等等。

3 现代汉语虚词用法自动识别

在自然语言处理领域中存在两种主流的研究方法，基于规则的研究方法和基于统计的研究方法。基于规则的理论主义方法在自然语言处理的研究初期占主流地位，关于该方法的研究也很多。从 20 世纪 90 年代开始，随着语料库构建困难的降低，基于统计方法的研究在自然语言处理领域越来越普遍，并且也取得了很多成果。近些年的研究发现，对自然语言处理的研究不能单独依靠基于规则的方法，也不能只研究基于统计的方法，应该将两者结合起来，采用规则和统计相结合的方法，并在情感分析[38]、中文命名实体识别[39]等领域已经取得了很好的效果。

虚词主要包含助词、语气词、连词、介词、方位词、副词等，在关于虚词用法的研究中，韩英杰[8]、周溢辉[9]、周丽娟[10]、袁应成[7]等分别对虚词中的助词、语气词、连词、介词等进行了用法自动识别研究。在副词用法的研究方面，刘锐等人分别采用基于规则[40]、基于统计[41]的方法对副词用法进行了自动识别的研究，本文主要在此基础上采用规则和统计相结合的方法对副词进行用法的自动识别。

3.1 基于规则的虚词用法自动识别研究

基于规则的方法主要以语言学的理论为基础，依靠语言学家等专家手工编写的规则来描述语言的一种语法，这种方法强调语言学家等专家对语言现象的认识。

采用袁应成[12]已经开发的基于规则的虚词用法自动标注系统（图 3.1）对现代汉语虚词的用法进行自动识别研究。

在虚词用法自动标注系统中进行用法规则的匹配时，系统是对规则进行逐条匹配的，系统默认用法编码信息在前的用法拥有较高的优先级，只要读取到一条规则与当前语境相匹配，便输出该规则的用法编码 ID，同时退出匹配程序，进行下一个词语的用法标注。

根据虚词用法词典转化的规则之间存在一定的包含或覆盖关系，一些规则在形式上和语法上还存在着交叉现象，这些规则本身上的不足可能导致一些用法的识别准确率很低。例如用法<d_dou1_1>的规则描述中仅要求副词“都”的后

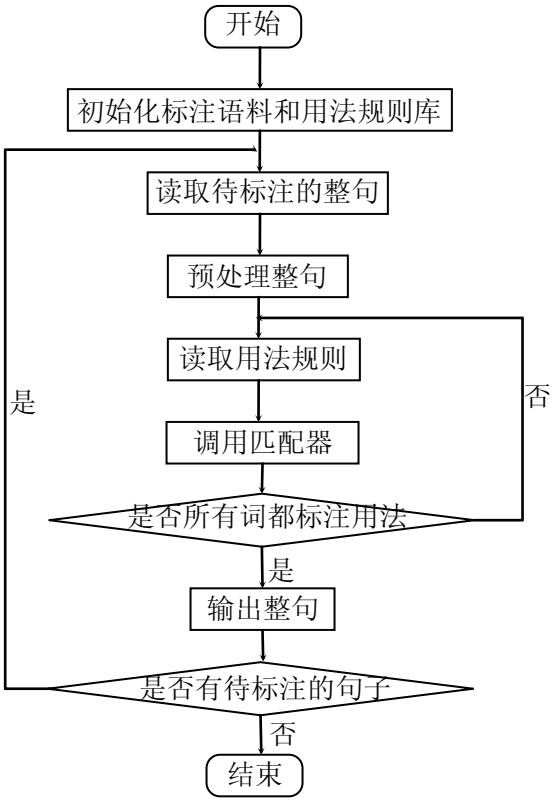


图 3.1 虚词用法自动标注系统流程图

面出现形容词或者动词，条件相对比较宽泛，用法<d_dou1_2>、用法<d_dou1_1b>等规则的描述条件也包含了或者暗含了该描述信息。如以下例句中副词“都”的标注结果：

不管/c “/wyz 台独/j ”/wyy 以/p 什么/ry 名义/n 、/wu 何种/r 方式/n 制造/v 分裂/vn 、/wu 谋求/v 独立/v ，/wd **都/d<d_dou1_1> 会/vu** 损害/v 两岸/n 人民/n 的/ud 根本/a 利益/n ，/wd **都/d<d_dou1_1> 会/vu** 使/v 祖国/n 和平/ad 统一/v 变/vi 得/ue 不/d 可能/v ，/wd 这/rz 是/vl 对/p 中国/ns 主权/n 和/c 领土/n 完整/a 的/ud 挑衅/vn 。/wj

该例句中副词“都”后面出现了动词，满足用法<d_dou1_1>中的规则描述“N→v”，而且用法<d_dou1_1>在规则中排在首位，拥有最高的优先级，所以利用虚词用法自动标注系统进行用法标注时将其标注为用法<d_dou1_1>，但是我们对该句子进行考察分析发现，副词“都”前面的语境中出现了词语“不管”，即

“M→不管”，另外结合上下文的意思可以判断此处满足用法<d_dou1_1b>的规则和词典描述，应该标注为用法<d_dou1_1b>。

为了解决上述问题，我们可以将用法的规则进行排序，重新调整词语各个用法的优先级（先后顺序），具体的步骤如下：

步骤 1：逐句读取《人民日报》分词和词性标注语料中的句子，直至语料文件末尾。在语料中查找虚词出现的位置，根据现代汉语虚词规则库中的规则描述，判断识别出词语在语境中的具体用法，即该词语用法的编码 ID，并将用法编码 ID 标注在词语的右边，形成系统标注用法的语料，即机标语料。

步骤 2：将机标语料中的虚词用法标注信息与虚词用法语料库中的标准答案进行对比，统计得出每个虚词用法自动识别的准确率、召回率等评价指标。

步骤 3：对统计得出的数据以及标注用法错误的语料进行考察分析，对虚词用法规则库中用法的规则进行调序、扩充、细化、整合等优化。

步骤 4：重复步骤 1~3，对虚词用法的规则不断进行优化分析，最终得到了使虚词用法自动识别准确率较高的规则库以及机标语料。

对用法规则进行优化调整后，副词“都”的用法规则描述为：

$都

@<d_dou1_1b>→M ^M→(不论|不管|无论|虽然|尽管|凡是|只要)*{，}

@<d_dou1_1d>→FR ^F→~ ^R→是

@<d_dou1_2a>→M ^M→连|甚至

@<d_dou1_2b>→A~A(不|没|没有|未|<df>) ^A→v

@<d_dou1_2c>→MN ^M→一 q ^N→不|没|没有|未|<df>

@<d_dou1_3>→E ^E→了，

@<d_dou1_2d>→N ^N→[，]*(不|没|<df>)

@<d_dou1_2b>→A~(不|没|没有|未|<df>)*A ^A→v

@<d_dou1_1c>→NE ^N→谁|哪里|什么|怎么|哪儿|哪|<ry>|<ryw> ^E→？

@<d_dou1_1a>→M ^M→谁|哪里|什么|怎么|哪儿|哪|<ry>|<ryw>

@<d_dou1_1>→N ^N→v|a

@<d_dou1_2>→N ^N→v

调整规则顺序后上述例句中副词“都”的用法标注如下：

不管/c “/wyz 台独/j ”/wyy 以/p 什么/ry 名义/n 、/wu 何种/r 方式/n 制造/v 分裂/vn 、/wu 谋求/v 独立/v ，/wd **都/d<d_dou1_1b>** 会

/vu 损害/v 两岸/n 人民/n 的/ud 根本/a 利益/n ，/wd **都/d<d_dou1_1b>** 会/vu 使/v 祖国/n 和平/ad 统一/v 变/vi 得/ue 不/d 可能/v ，/wd 这/rz 是/vl 对/p 中国/ns 主权/n 和/c 领土/n 完整/a 的/ud 挑衅/vn 。/wj

本文所有的实验都采用 2000 年 4~6 月《人民日报》分词和词性标注后的语料。副词“都”在现代汉语中出现的频率比较高，其用法和意义也比较复杂，对其进行单独的研究和分析也有助于其他副词甚至虚词的用法自动识别研究。

以副词“都”为例对每个用法的标注情况进行分析，副词“都”在语料中共出现 6791 次，各个用法在语料中的分布率以及词频如表 3.1 所示。由表 3.1 可以看出，因为《人民日报》每天都会有新的内容，又 4 月和 6 月为 31 天，而 5 月为 30 天，所以 5 月的副词“都”词频较少，除此之外三个月的语料差别不大，而且副词“都”的每个用法在三个月的语料中的分布率相近。因此，本文在进行实验研究分析时，可以排除语料分布不均匀而影响实验结果的因素。

表 3. 1 副词“都”的用法分布

用法	2000 年 4 月		2000 年 5 月		2000 年 6 月		三个月总计	
	分布率(%)	词频	分布率(%)	词频	分布率(%)	词频	分布率(%)	词频
d_dou1_1	84.18	1963	80.92	1650	86.20	2086	83.9	5699
d_dou1_1a	2.53	59	3.63	74	1.94	47	2.7	180
d_dou1_1b	9.31	217	9.17	187	8.55	207	9.00	611
d_dou1_1c	0	0	0.20	4	0.21	5	0.13	9
d_dou1_1d	0.04	1	0.34	7	0	0	0.12	8
d_dou1_2	0.64	15	1.23	25	0.12	3	0.60	43
d_dou1_2a	2.06	48	3.24	66	1.90	46	2.40	160
d_dou1_2b	0.13	3	0	0	0.04	1	0.06	4
d_dou1_2c	0.17	4	0.29	6	0.17	4	0.21	14
d_dou1_2d	0.30	7	0.25	5	0	0	0.18	12
d_dou1_3	0.64	15	0.74	15	0.87	21	0.75	51
合计	100.00	2332	100.00	2039	100.00	2420	100.00	6791

利用优化后的规则库对实验语料中的副词“都”进行用法自动识别，可以得到基于规则方法的用法标注语料，与语料库中的正确标注进行对比，结果如表 3.2 所示。在语料中出现的 6791 个副词“都”中，采用基于规则的虚词用法自动标注系统正确标注的有 5555 个，规则标注的准确率为 82%。

虽然副词“都”各个用法的分布率有很大的差异，但对各个用法的分布率和

识别情况进行分析发现，用法的分布率对用法的识别情况没有明显影响，如分布率较高的用法<d_dou1_1>和分布率较低的用法<d_dou1_1b>，其识别效果都较好。

表 3.2 基于规则的副词“都”的用法自动识别

用法	分布率(%)	准确率(%)	召回率(%)	F 值(%)
d_dou1_1	83.9	99.61	80.19	88.85
d_dou1_1a	2.7	74.46	76.11	75.27
d_dou1_1b	9.00	89.40	99.35	94.11
d_dou1_1c	0.13	100.00	77.77	87.50
d_dou1_1d	0.12	3.64	100.00	7.02
d_dou1_2	0.60	0	0	0
d_dou1_2a	2.40	85.25	97.50	90.96
d_dou1_2b	0.06	100.00	50.00	66.67
d_dou1_2c	0.21	61.11	78.57	68.75
d_dou1_2d	0.18	1.09	75.00	2.15
d_dou1_3	0.75	68.57	94.12	79.34

以“别”、“不”、“才”等十个副词为例对每个词语的标注情况进行分析（表3.3），每个词语出现的频率有很大的差异，并且其用法的识别情况差别也较大，但两者没有直接的关系。

表 3.3 基于规则方法的副词用法自动识别

词语	词频	正确数	准确率(%)	词语	词频	正确数	准确率(%)
别	140	122	87.14	倒	53	15	28.30
不	16355	15278	93.41	倒是	35	17	48.57
才	2029	914	45.05	多	666	98	14.71
大概	33	22	66.67	分别	895	455	50.84
当然	363	275	75.76	更	3745	3599	96.10

除去用法的分布率以及词语出现频次两个因素的影响，发现影响基于规则方法识别效果的是规则本身。在规则的生成和优化过程中，需要对大量的文献进行详细分析，对语料不断地进行考察、分析和实验。同时也需要语言学家等各种专家的配合，研究工作本身耗时耗力，但毕竟是人为的工作，不能保证完全不参杂主观因素，这些会影响对虚词用法的客观描述。

另外，在现代汉语虚词用法词典的用法描述中，包含着很多在现代汉语虚词用法规则中无法体现出来的信息特征，如读音、语气、情感倾向、短语等。这些问题直接导致规则不能完全表达用法的描述信息，也影响了用法自动识别的效果。如以下例句：

这些/rz 人/n 互相/d 见{jian4}/v 了/ul 面/n ，/wd 打招呼/vi 没{mei2}/v 一个/mq 真/a 名/Ng ，/wd **都/d<d_dou1_1d> 是/vl** 别号/n 绰号/n ，/wd 甚或/c 还有/v 人/n 叫/vl “/wyz 座山雕/nr ”/wyy 、/wu “/wyz 胡/nrf 汉三/nrg ”/wyy 的/ud ，/wd 这/rz 当然/d 是/vl 取/v 其/rz 形似/v 而/c 非/Vg 神似/a 了/y 。/wj

该例句中副词“都”在小句中位于句首（即 F→~），且后面紧邻着“是”（R→是），满足用法<d_dou1_1d>的规则描述，因此在进行基于规则方法的自动识别时，将其标注为用法<d_dou1_1d>。对比副词“都”的规则以及词典描述，分析该例句表达的意思，发现在本句中，副词“都”表示总括全部，即所有打招呼时用的称呼，虽然与“是”搭配，但不是说明原因，也没有责备的意思，所以此句中的用法标注应该为用法<d_dou1_1>。对标注错误的原因进行分析，发现这种错误主要是因为用法<d_dou1_1d>的规则描述中缺少了感情色彩“责备”，没有完全表达词典中用法的释义。

3.2 基于统计的虚词用法自动识别研究

基于统计的经验主义研究方法主要采用统计学处理技术从大规模的语料库中半自动或者自动地获取研究所需要的语言信息，并建立有效的统计模型，根据训练数据的实际情况不断地进行优化，而上节中讨论的基于规则的理论主义的研究方法则很难做到这一点。

在上节的讨论中，我们发现基于规则的方法在虚词的用法自动识别中已经得到了很好的效果，但同时也存在一些很难解决的困难，本节将分析讨论基于统计的虚词用法自动识别研究。

在自然语言处理领域，有许多机器学习的统计模型，如：隐马尔科夫模型（Hidden Markov Model, HMM）、最大熵（Maximum Entropy, ME）模型、支持向量机（Support Vector Machine, SVM）模型、决策树（Decision Tree）、条件随机场（Conditional Random Fields, CRF）模型等等，这些统计模型在自然语言处

理领域中都得到了广泛的应用。考虑到虚词的用法与虚词所在的上下文信息及语境序列有着密切的限制依赖关系，本文选自了条件随机场（CRF）模型、最大熵（ME）模型、支持向量机（SVM）模型对虚词用法自动识别进行分析研究，这三个模型在序列标注方面效果比较好。

3.2.1 条件随机场模型

2001 年 Lafferty 等人[14]首先将条件随机场模型引入到了自然语言处理的序列标注和切分有序数据的学习任务中。条件随机场模型是一个无向图模型，该模型在给定的输入节点序列的条件下，计算输出节点的条件概率，它考察的是输入节点序列对应的标注序列的条件概率，训练目标是使条件概率值最大。其核心思想主要是利用无向图的理论提高序列标注和切分有效数据的结果准确率等指标。

CRF 模型最早是用于解决序列数据分析的问题，现在已经应用于生物信息学、机器视觉、自然语言处理[42] [43] [44]以及网络智能等领域，并取得了很好的效果。现代汉语虚词用法自动识别也可以看做为一个序列标注问题，根据虚词所在的上下文语言环境信息确定虚词的用法。

本文在基于 CRF 的虚词用法自动识别研究的实验中，利用 CRF 模型的实现工具包 CRF++0.53*对模型进行训练和测试。

在采用条件随机场模型进行研究的过程中，特征的选取是一个很重要的环节，特征选取的好坏直接影响实验结果的好坏。换言之，进行特征选取时，我们要尽量选取那些对虚词用法识别有利的因素。

通过对语料中虚词用法的大量分析，我们发现虚词上下文中的词语和词性信息对该虚词的用法识别有很大的影响。因此在采用条件随机场模型对现代汉语虚词用法进行自动识别研究时，选取虚词上下文中词语信息、词性信息以及词语与词性的复合信息作为 CRF 的特征模版，通过调整上下文的窗口大小选取不同的特征进行实验，其中选取上文词语 m 个，下文词语 n 个，m、n 分别控制在 2 到 10。

在利用工具包 CRF++0.53 进行实验之前，需要将已有的语料按照上述特征模版转换为工具包可识别的格式。以虚词中的副词“都”为例子，m 和 n 分别选取 3 和 2，下述例句转换后的格式如表 3.4 所示。

*http://crfpp.softwareforce.net

他/rr 说/v ，/wd **中国/ns 和/c 秘鲁/ns 都/d<d_dou1_1> 拥有/v 灿烂/a** 悠久/a 的/ud 文化/n 传统/n ，/wd 两/m 国/n 同/d 属/vl 发展中国家/l ，/wd 对/p 许多/m 重大/a 国际/n 问题/n 持有/v 相同/a 或/c 相近/a 的/ud 看法/n ，/wd 进一步/d 加强/v 双方/n 在/p 各个/rz 领域/n 的/ud 友好/a 合作/vn ，/wd 潜力/n 巨大/a ，/wd 前景/n 广阔/a 。/wj

表 3. 4 CRF 模型中副词“都”的数据转换示例

0	1	2	3	4	5	6	7	8	9	10	11	12
都	d	中国	ns	和	c	秘鲁	ns	拥有	v	灿烂	a	<d_dou1_1>

实验时选择现代汉语虚词词用法语料库中《人民日报》2000 年 4~6 月的语料进行十折交叉实验，副词“都”在不同的窗口上的实验结果如表 3.5 所示。

表 3. 5 基于 CRF 模型不同窗口的实验结果

窗口	2	3	4	5	6	7	8	9	10
2	88.25	87.44	87.48	87.26	87.26	86.66	86.49	86.66	86.28
3	87.26	87.26	87.14	87.26	86.75	86.36	86.19	86.19	86.45
4	87.91	87.56	87.26	87.22	86.96	86.88	86.45	86.49	86.79
5	88.51	87.48	87.52	87.22	87.39	87.14	87.22	87.09	86.96
6	88.96	87.82	87.95	87.69	87.65	87.22	86.79	86.79	86.79
7	89.43	89.62	88.59	87.99	87.44	87.35	86.92	86.88	86.84
8	88.72	88.04	87.91	87.39	87.22	87.26	86.96	86.96	86.96
9	89.07	88.04	87.56	87.65	87.48	87.14	87.05	87.14	87.01
10	88.98	88.72	88.38	88.42	87.61	87.44	87.44	87.18	86.71

由表 3.5 可以看出，选择左窗口为 7 右窗口为 3 时，实验的效果最好，副词“都”各用法自动识别情况如表 3.6 所示。在语料中出现的 6791 个副词“都”中，正确标注的有 6086 个，准确率为 89.62%。

由表 3.6 可以看出，与基于规则的实验结果不同，副词“都”各个用法的分布率对各用法的识别效果有很大的影响，分布率高的用法识别准确率也较高，反之亦然。对比基于规则与基于 CRF 统计方法的结果，我们可以发现，虽然基于 CRF 的统计方法在总体上优于基于规则方法的识别结果，但是对于某些用法来说，采用基于 CRF 的统计方法识别效果比较差，例如用法<d_dou1_1a>、用法<d_dou1_2a>、用法<d_dou1_2c>等等，这主要是因为这些用法在语料中的分布较少，在进行统计学习时不容易学习到。

表 3.6 基于 CRF 的副词“都”各用法识别情况

用法	准确率(%)	召回率(%)	F 值(%)
d_dou1_1	89.84	99.28	94.32
d_dou1_1a	78.81	51.67	62.42
d_dou1_1b	86.92	36.99	51.90
d_dou1_1c	0	0	0
d_dou1_1d	0	0	0
d_dou1_2	50.00	2.33	4.45
d_dou1_2a	100.00	65.63	79.25
d_dou1_2b	0	0	0
d_dou1_2c	0	0	0
d_dou1_2d	0	0	0
d_dou1_3	42.86	5.88	10.34

采用同样的方法对副词进行实验，实验结果如表 3.7 所示。不同的词语有不同的最佳识别窗口，识别的效果也不同，需要对每个词语都进行实验分析。此外，针对副词而言，采用 CRF 统计模型进行实验时其用法与其距离较近的上下文信息关系密切。

表 3.7 基于 CRF 方法的副词用法识别情况

词语	左窗口	右窗口	准确率(%)	词语	左窗口	右窗口	准确率(%)
别	2	5	86.96	倒	3	3	62.26
不	2	4	96.93	倒是	2	5	51.43
才	3	4	94.53	多	4	3	89.34
大概	2	2	60.87	分别	5	4	62.13
当然	4	3	83.65	更	2	2	94.28

3.2.2 最大熵模型

“熵”最初是一个热力学概念，20 世纪 40 年代香农首先在信息论中加入了信息熵的概念。最大熵方法最先是由 E.T.Jaynes 在 1975 年提出来的[15]，后来的学者提出了与之对应的算法来估计统计模型的参数。由于各种条件的限制，最大熵模型在提出后的初期，其本身的优势并没有在自然语言处理领域得到很好的体现。直到 20 世纪 90 年代，最大熵模型的框架和算法得到了描述和实现，并应用到了自然语言处理领域的研究任务中，同时也取得了很好的效果，之后便引起了研究者的广泛关注，将其广泛应用于自然语言处理的各个领域，而且都取得了成功。

最大熵模型的基本思想是，对已知的内容进行建模，对未知的内容不做任何假设，没有任何偏见。即承认已知的事物和信息，在建模的时候应该尽量符合已知的信息，而对于未知部分使模型的熵最大（不确定性最大）。其实质就是，在已知的部分知识的前提之下，关于未知部分的最合理的推断便是符合已知条件的最随机也是最不确定的推断，也就是我们可以做出的最佳选择。基于 ME 建立的语言模型不依赖于领域知识，独立于特定的任务，适合自然语言处理和自然语言理解各个领域的应用和研究，在词性标注、词义消歧、命名实体识别、文本分类等领域已经取得了很好的效果。

本文在基于 ME 的虚词用法自动识别研究的实验中，利用 ME 模型的工具包 maxent[†]对模型进行训练和测试实验。

与采用条件随机场模型进行研究相似，采用最大熵模型进行研究时也需要选取特征。本文同样选取虚词上下文中的词语信息、词性信息以及词语和词性的复合信息作为 ME 的特征模板，不同的是，为了减少实验的次数，我们将上下文的窗口选取相同的值 k，且将 k 的值控制在 2 到 10 之间。实验时同样选择现代汉语虚词用法语料库中 2000 年 4~6 月的语料进行实验，实验结果如表 3.8 所示，与 CRF 模型类似，每个词语的最佳窗口都不同。从总体上来看，采用 ME 模型的实验结果略微优于 CRF 模型的实验效果。

表 3. 8 基于 ME 方法的副词用法识别情况

词语	窗口	准确率(%)	词语	窗口	准确率(%)
都	9	88.87	倒	3	66.80
别	3	88.45	倒是	2	30.00
不	2	97.23	多	2	84.48
才	3	83.98	分别	8	62.44
大概	2	53.33	更	9	93.90
当然	8	77.47			

最大熵模型有很多优点，在建模的时候，只需要进行特征的选取，而无需考虑如何使用这些特征；选取特征的过程也很灵活，进行实验的时候也可以灵活的更换特征得到不同的实验结果，进而方便对实验结果进行分析总结；该模型可移植性强，可以应用到不同的领域中；可结合丰富的信息。

当然，最大熵模型也有自己的缺点不足，利用最大熵模型进行实验时，对时间和空间的要求都较高，时空开销较大；数据的稀疏问题也比较严重，存在标

[†] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

注的偏置问题；进行特征选择后也无法对特征进行融合；另外，进行最大熵模型实验时，对语料库的依赖性较强，实验所用的语料对实验的结果影响较大。

3.2.3 支持向量机模型

支持向量机（Support Vector Machine, SVM）是由 Vapnik[16]提出来的一种统计学习方法，SVM 在解决高维模式识别、非线性以及小样本的任务中体现出了许多其自身特有的优势，其中在分类方面有良好的性能，并且可以推广应用到其他的机器学习问题当中。近些年，支持向量机模型的应用得到了研究者的广泛关注和应用，不仅应用到知识发现、模式识别等领域的理论研究中，还涉及了自然语言处理、生物学以及图像处理等相关技术领域的应用研究，并且都取得了很好的效果。其中，在自然语言处理领域的研究中，支持向量机主要应用在词义消歧、短语识别、信息过滤和文本分类等方面。

本文在基于 SVM 的虚词用法自动识别研究的实验中，利用 SVM 模型的工具包 LibSVM‡对模型进行训练和测试实验。

与条件随机场模型和最大熵模型不同，支持向量机模型的特征是数值型的。因此首先要定义该值的计算方法。对于句子中出现的虚词，通过选定该词所在的上下文的窗口，然后计算窗口内的特征词语与该词用法的互信息，以此来作为特征向量。

实验时将上下文的窗口选取相同的值 m，且将 m 的值控制在 2 到 10 之间。实验时同样选择现代汉语虚词用法语料库中 2000 年 4~6 月的语料进行实验，，实验结果如表 3.9 所示，每个词语的最佳窗口都不同，从总体上来看，采用 SVM 模型进行实验的结果明显优于 CRF 模型和 ME 模型的实验结果，但是以花费大量时间为代价的。

表 3.9 基于 SVM 方法的副词用法识别情况

词语	窗口	准确率(%)	词语	窗口	准确率(%)
都	2	96.14	倒	4	72.43
别	9	94.62	倒是	6	63.59
不	4	98.27	多	10	68.59
才	2	97.40	分别	5	65.91
大概	6	94.83	更	2	97.24
当然	5	91.53			

‡ http://www.csie.ntu.edu.tw/~cjlin/libsvm

三种统计模型对虚词用法的自动识别都有很好的效果（图 3.2），其中采用 SVM 模型效果最好，其次是 CRF 模型，而 ME 模型的效果比 CRF 略微差一点。从时间复杂度角度分析，发现采用 SVM 模型进行实验需要花费大量的时间，以时间为代价来得到更高的识别效果，而其他两种模型的时间复杂度较低。

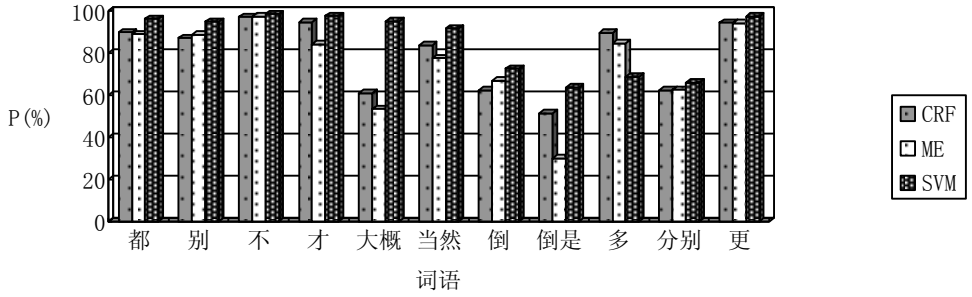


图 3.2 三种统计模型结果对比

3.3 规则与统计相结合的虚词用法自动识别研究

一般认为，在自然语言处理领域中存在两种主流的研究方法，基于规则的理性主义研究方法和基于统计的经验主义研究方法，但是，这两种方法并非是对立的。在实际的研究中，我们不能单纯的依靠规则方法，也不能只依靠统计方法，应该分析这两种方法各自的优点和不足，将理性主义和经验主义紧密的结合起来。

针对本文中关于现代汉语虚词用法自动识别的研究，基于规则的方法和基于统计的方法都有各自的优点，为了取得更好的用法自动识别结果，我们考虑采用规则与统计相结合的方法，综合考虑规则和统计的优点，最终实现用法自动识别准确率的提高。

我们在对基于统计和基于规则这两种方法的实验结果进行分析时，都需要对每个用法的分布率、召回率、准确率等参数进行分析考察，因此我们考虑这些参数信息在规则和统计实验中的相互关系，并将其加入到规则和统计相结合用法自动识别的过程中。例如，采用基于规则的方法时，分布率低的用法识别情况比较好，可以将这些用法的规则信息或者识别的结果作为前提参考信息加入到统计学习的过程中；在进行统计方法学习之前，先识别出分布率较低的用法，然后发挥统计方法的优势，识别出分布率较高的用法。经过分析考察，得到了

以下思路方法。

方法一：针对每个虚词用法分布以及识别的情况，为每个虚词设定一个显示分布率高低的阈值。如果用法的分布率低于该阈值就将其归为分布率较低的用法，反之属于分布率较高的用法。对虚词用法进行自动识别时，先进行规则方法的识别得到用法标注结果，如果该用法的分布率低于设定的阈值，就直接输出该结果；反之放弃该结果，转而采用统计的方法进行识别标注。

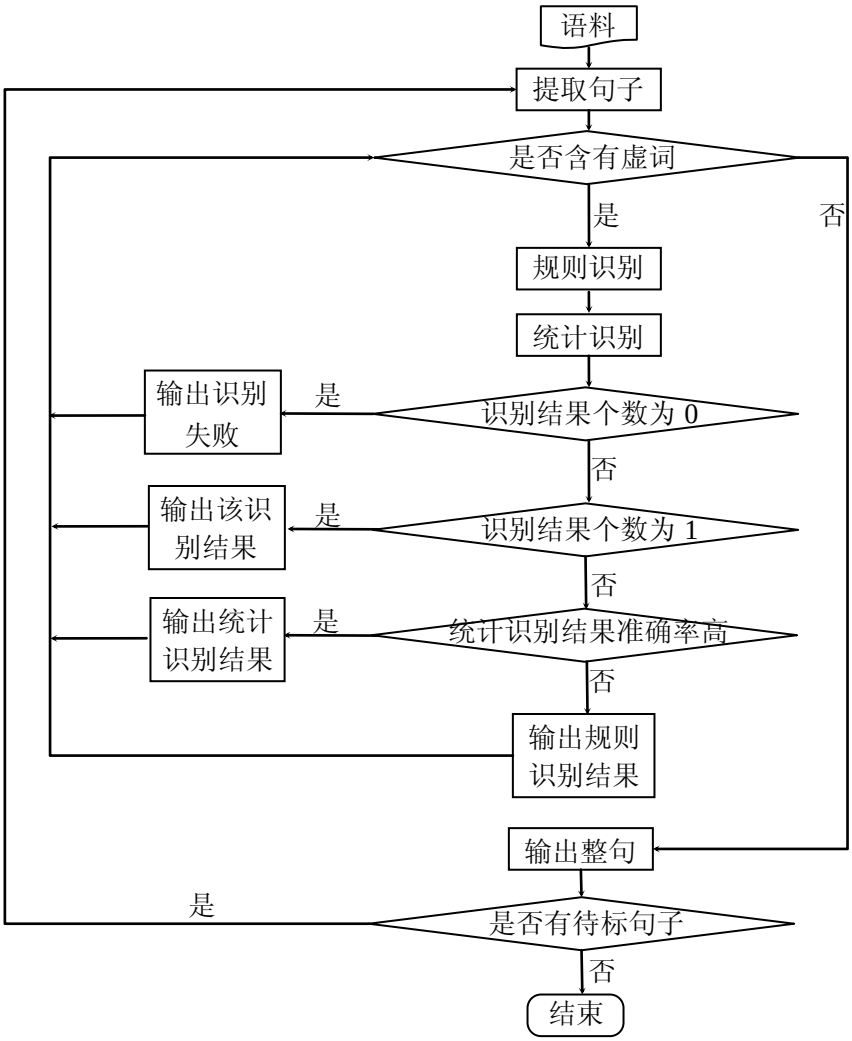


图 3.3 规则与统计相结合方法的流程图

方法二：对虚词分别进行基于规则的方法和基于统计的方法自动识别，识别结果分为以下三种情况：

a) 识别出零个结果：即采用两种方法都未识别出结果，我们认为该虚

词在上下文语境的影响下用法识别失败；

b) 识别出一个结果：或者规则方法没有识别结果，或者统计方法没有识别结果，或者两种方法的识别结果相同，此时输出该结果；

c) 识别出两个结果：即两种方法都能识别且识别结果不同，此时我们对比两个结果在各自单独进行识别时的准确率大小，将准确率高的结果作为最终的识别结果。

本文在进行规则与统计相结合的虚词用法自动识别研究时，基于统计的方法选择CRF模型，以副词“都”为例，采用方法二对其进行用法自动识别的流程如图3.3所示。

对实验语料进行规则与统计相结合的实验，副词“都”的用法识别结果如表3.10所示，识别的准确率达到98.54%，高于基于规则的方法和基于统计的方法的识别准确率。从用法识别的角度分析，除了分布率高的用法<d_dou1_1>具有较好的识别效果以外，分布率低的用法识别效果与CRF模型相比也有很大提高，如用法<d_dou1_3>等。

表 3. 10 规则与统计相结合的副词“都”用法自动识别

用法	准确率(%)	召回率(%)	F 值(%)
d_dou1_1	99.58	99.65	99.62
d_dou1_1a	97.77	97.22	97.49
d_dou1_1b	96.45	97.87	97.15
d_dou1_1c	77.78	77.78	77.78
d_dou1_1d	87.50	87.50	87.50
d_dou1_2	7.41	4.65	5.71
d_dou1_2a	99.38	99.38	99.38
d_dou1_2b	100.00	50.00	66.67
d_dou1_2c	73.33	78.57	75.86
d_dou1_2d	44.44	66.67	53.33
d_dou1_3	88.00	86.27	87.13

采用同样的方法对其它副词进行实验，实验结果如表 3.11 所示。

基于规则方法、基于统计方法以及规则和统计相结合方法的实验结果对比如图3.4所示，总体上来看，采用规则与统计相结合方法进行虚词用法自动识别的识别效果有所提高，但这种相结合的思路并不使用于每一个虚词，对于规则结果和统计结果差别较大的词语（如“多”）来说，相结合的方法可能会略差与基于统计方法的结果。

表 3.11 规则与统计相结合的副词用法识别情况

词语	准确率(%)	词语	准确率(%)
别	89.29	倒	56.41
不	97.12	倒是	53.18
才	89.75	多	68.83
大概	72.73	分别	65.94
当然	85.21	更	97.51

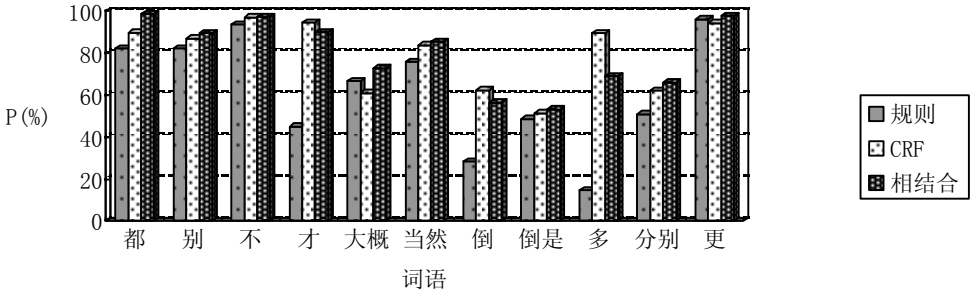


图 3.4 规则、统计以及相结合实验结果对比

3.4 本章小结

本章分别采用三种方法对现代汉语虚词用法进行自动识别研究。在基于规则的用法自动识别中，根据虚词的规则通过标注系统对语料进行标注分析，并不断的优化虚词的规则描述，以便提高规则方法的效果。基于统计的方法研究中，分别介绍了条件随机场模型、最大熵模型、支持向量机模型，并采用这三种统计模型进行实验分析。最后将基于规则的方法和基于统计（CRF）的方法进行结合，实验结果表明，采用规则和统计相结合的方法可以提高虚词用法自动识别的效果。

4 虚词用法识别在依存句法分析中的应用

现代汉语中的虚词是汉语句子的重要组成部分，对理解句子有重要的影响。以上章节主要对虚词的用法进行自动识别研究，其目的是更准确的理解虚词的意思，理解整个句子的意思，这涉及到虚词用法的一个应用领域，也是中文信息处理的一个重点难点——句法分析。句法分析主要有两种思路，短语结构句法分析和依存句法分析。近些年，依存句法分析逐渐成为研究的热点，其中关于汉语的依存句法分析也取得了很好的成绩[45][46]，但是目前现有的存句法分析器很少考虑到句子中虚词用法的影响，本章主要考虑在虚词用法自动识别的基础上将虚词的用法运用到汉语依存句法分析中，试图提高依存句法分析的效果。

4.1 LTP 中的依存句法分析

对句法分析而言，树库是一个非常重要的资源，树库的质量和规模直接影响句法分析的性能。当前最有名的树库美国宾夕法尼亚大学构建的 Penn Treebank，为句法分析的研究者提供了一个平台，同时也促进了句法分析的研究发展。在汉语方面构建的树库多数是采用短语结构的标注形式，如清华大学的汉语树库、美国宾夕法尼亚大学 Penn Chinese Treebank (Penn CTB)树库，以及台湾中央研究院的 Sinica 树库等。目前汉语依存树库的建设还有很多的不足，在规模和质量上与英语等其他语言还有很大的差距，这些阻碍了汉语依存句法分析的研究和发展。本文在对依存句法进行分析时，使用了哈尔滨工业大学社会计算与信息检索研究中心研制的语言技术平台（LTP）[30]以及依存树库 HIT-IR-CDT[47]。

LTP 提供了一整套的自底向上的中文语言处理模块，包括分词、词性标注、命名实体识别、依存句法分析、词义消歧、语义角色标注等六项中文处理技术，高层的处理技术都是在低层技术的基础上进行的，其中依存句法分析依赖分词和词性标注的结果，词语之间相互关系采用 24 种依存关系来表示。

通过详细的分析对比发现，分词以及词性标注的结果对依存句法分析影响很大，而本文着重研究虚词的用法对依存句法分析的影响，所以对标准的分词和词性标注结果进行依存句法分析，减少因分词和词性标注错误对依存句法分析的影响，进而方便对依存句法分析结果的分析和研究。

首先对 HIT-IR-CDT 中的 1000 句测试实例进行依存句法分析，24 种依存关系的识别结果如表 4.1 所示, 其中“-R”表示依存关系的父节点识别正确，但是依存关系识别错误的个数，“-PR”表示父节点识别错误的个数，LAS 表示依存关系的识别准确率，UAS 表示依存弧的识别准确率。

表 4. 1 24 种依存关系的识别情况

关系类型	符号	结点数	-R	-PR	UAS(%)	LAS(%)
定中关系	ATT	4755	11	174	96.34	96.11
状中结构	ADV	2704	43	64	97.63	89.39
动宾关系	VOB	2358	62	66	97.20	94.57
主谓关系	SBV	1613	42	48	97.02	94.42
“的”字结构	DE	1081	1	41	96.21	96.11
核心	HED	1000	0	19	98.10	98.10
介宾关系	POB	817	4	21	97.43	96.94
连动结构	VV	730	141	26	96.44	77.12
并列关系	COO	701	54	58	91.72	84.02
数量关系	QUN	618	5	4	99.35	98.54
独立分句	IC	334	110	28	91.62	58.68
语态结构	MT	308	0	14	95.45	95.45
前附加关系	LAD	262	2	8	96.95	96.18
动补结构	CMP	233	16	5	97.85	90.99
关联结构	CNJ	226	6	3	98.67	91.59
后附加关系	RAD	111	0	3	97.30	97.30
独立结构	IS	89	14	7	92.13	76.40
同位关系	APP	77	14	7	90.91	72.73
“地”字结构	DI	33	0	1	96.97	96.97
依存分句	DC	20	15	1	95.00	20.00
“得”字结构	DEI	12	0	0	100.00	100.00
比拟关系	SIM	4	0	1	75.00	75.00
“把”字结构	BA	0	0	0	——	——
“被”字结构	BEI	0	0	0	——	——

24 种依存关系中，定中关系（ATT）、动宾关系（POB）、主谓关系（SBV）等依存关系的识别效果较好，连动结构（VV）、并列关系（COO）、独立分句（IC）、独立结构（IS）、同位关系（APP）等依存关系的识别效果较差。

如果 LAS 和 UAS 两个指标的差别较大，则该依存关系的识别问题主要出在依存关系的确定上，即“-R”的值远高于“-PR”，如依存分句（DC）；倘若两个

指标的值基本相同，则问题主要在父节点的识别错误上，如“DE”字结构（DE），定中关系（ATT）。对定中关系 ATT 的错误情况进行分析，发现当中心词前出现多个修饰语时，修饰语之间以及修饰语与中心语之间的依存关系容易识别错误，如以下例句：

（1）**社会主义现代化建设**需要数以亿计高素质的劳动者和数以千万计的专门人才，除了思想和专业方面的要求外，还应当使他们具有较高的语言文字能力。

该例句在 LTP 中依存句法分析的部分结果如图 4.1 a 所示，名词“建设”在句子中作主语，“现代化”是“建设”的定语，“社会主义”又是“现代化”的定语。分析发现，“社会主义建设”和“现代化建设”更符合语言表达，即“社会主义”和“现代化”同时作为定语修饰“建设”，标准的句法分析结果为图 4.1 b。

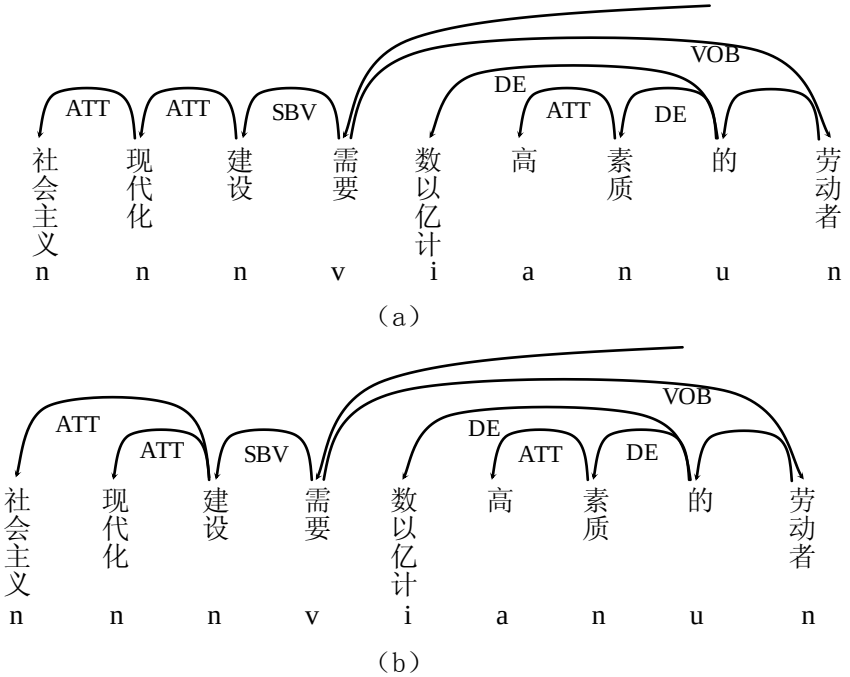


图 4. 1 例句（1）的句法分析结果

定中关系主要涉及名词、代词等实词，与虚词及其用法没有明显的联系。而在识别效果相对较差的依存关系中，并列关系 COO 一般与连词一起出现，与虚词的关系比较密切，因此主要分析虚词中连词及其用法在并列关系 COO 识别中的应用。

4.2 连词结构短语在依存句法分析中的应用

并列关系的句子中大都包含了诸如“和”、“并”、“及”、“或者”等连词标记，并且并列成分与这些连词结合共同构成连词结构短语，因此，可以根据连词用法识别出连词结构短语，然后确定其中的并列成分，最终识别并列关系。

4.2.1 并列关系的识别情况

在 1000 条测试语料中，并列关系共出现的 701 次，父节点识别错误的有 58 个，父节点识别正确而依存关系识别错误的有 54 个，对并列关系识别的错误类型进行总结（表 4.2），其中“COO-R-”表示并列关系的父节点识别正确，但依存关系识别错误；“COO-PR-”表示父节点识别错误，例如“COO-PR-ATT”表示某一结点因父节点识别错误而将依存关系误标为定中关系 ATT。

表 4.2 并列关系识别错误分类

错误类型	错误依存关系	个数	总计
COO-R-	VV	39	54
	IC	6	
	APP	2	
	VOB	6	
	CMP	1	
COO-PR-	VOB	9	58
	ATT	18	
	COO	9	
	VV	7	
	DE	6	
	SBV	5	
	ADV	2	
	POB	1	
	HED	1	

由表 4.2 可以看出，并列关系识别错误的主要原因有以下两个：

a）当两个或两个以上的动词并列连用时，易识别为连动结构，如以下例句：

（2）经过电话及时指导，便可立即**排除**故障或**进入**正常操作，减少处理过程和维修的盲目性，提高了医疗设备的使用率。

该句在 LTP 中依存句法分析的部分结果如图 4.2a 所示，句子中“排除”和“进入”虽都为动词，但由“或”连接，两者属于选择结构，并非两个动作的承接，

因此，两者是并列关系，标准的依存句法分析结果如图 4.2b 所示。

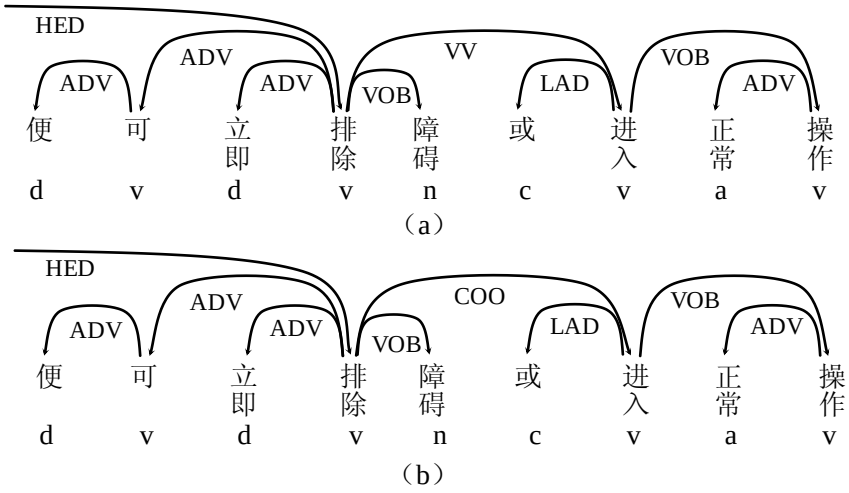


图 4. 2 例句（2）的句法分析结果

b）当复合短语中的修饰部分包含并列关系时，易识别为并列成分对核心词的修饰关系，如以下例句：

（3）这个举动，体现了该站抓好下岗职工基本生活**保障和再就业工作**的决心。

该句在 LTP 中依存句法分析的部分结果如图 4.3a 所示，句子中的“下岗职工基本生活保障”和“再就业”是构成并列关系的两个并列成分，虽然识别出了并列关系，但是两个并列成分识别错误，标准的依存句法分析结果如图 4.3b 所示。

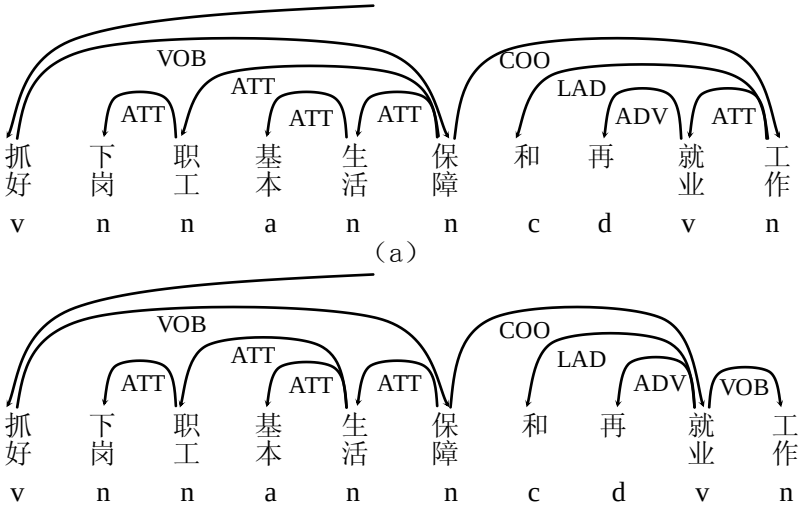


图 4. 3 例句（3）的句法分析结果

经过对依存树库中包含并列关系句子的分析，发现了包含并列关系的句子中大都包含了诸如“和”、“并”、“及”、“或者”等连词标记，或者含有顿号等标点符号标记。另外，并列成分之间的依存关系满足左核心原则，即由多个顿号连接的并列成分时，后面并列成分依存于第一个并列成分，且并列标记依存于右侧的并列成分。如下面的句子：

（4）徐先生还具体帮助他确定了把画雄鹰、鳜鱼、斑鸠、松鼠、麻雀和竹、梅、松、柏作为主攻目标。

其句法分析部分结果图 4.4 所示，句子中包含了连词标记“和”以及顿号标记“、”，并列成分“雄鹰”、“鳜鱼”、“斑鸠”、“松鼠”和“麻雀”构成一个并列结构，并列成分“竹”、“梅”、“松”和“柏”构成一个并列结构，另外“雄鹰”和“竹”作为两个并列结构的中心词又组成一个并列结构。

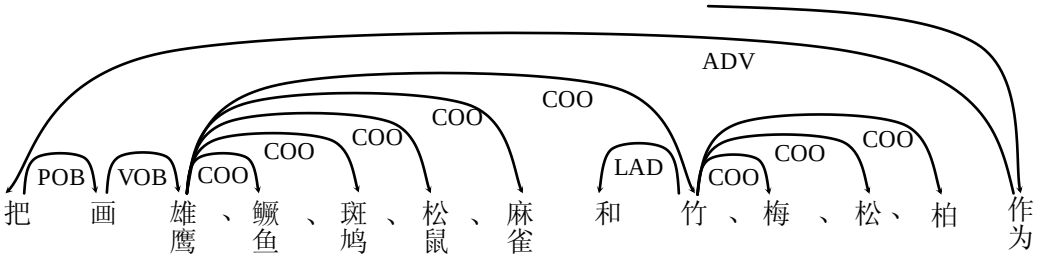


图 4.4 例句（4）的依存句法分析结果

在实际的文本中，并非所有的并列关系都包含上述并列标记，也并非都满足上述原则，另外还有上下文语境的影响，这些都影响了并列关系的识别。

综合以上的分析，考虑将与并列关系相关的连词的用法信息加入并列关系的识别中，主要包括表示并列结构的连词“和”、“与”、“及”、“并”等，以及表示选择结构的连词“或”、“或者”等，本文主要分析这些连词对并列关系识别的影响。

4.2.2 连词用法在并列关系识别中的应用

与并列关系相关的连词主要有“和”、“与”、“并”、“还是”、“或者”、“及”、“或”等等，这些连词与并列成分一起构成并列结构短语或者选择结构短语，所以采用周丽娟[10][48]对连词的用法及连词结构短语识别的相关研究，对句法分析的结果进行优化，提高并列关系的识别效果。

为了清楚的分析连词及其用法对并列关系识别的影响，以例句（5）为例进

行详细的分析：

（5）认为李自成占领北京后，中国面临的是**统一还是分裂**问题。

步骤一：从 HIT-IR-CDT 中抽出包含并列关系信息连词的句子，其中包含了句子的分词、词性标注、句法分析结果等信息，将其作为实验的对比语料。例句（5）包含的信息如下：

<sent id="9437" cont="认为李自成占领北京后，中国面临的是统一还是分裂问题。">

<word id="0" cont="认为" pos="v" parent="-1" relate="HED" />

<word id="1" cont="李自成" pos="nh" parent="2" relate="SBV" />

<word id="2" cont="占领" pos="v" parent="4" relate="ATT" />

<word id="3" cont="北京" pos="ns" parent="2" relate="VOB" />

<word id="4" cont="后" pos="nd" parent="7" relate="ADV" />

<word id="5" cont="，" pos="wp" parent="-2" relate="PUN" />

<word id="6" cont="中国" pos="ns" parent="7" relate="SBV" />

<word id="7" cont="面临" pos="v" parent="8" relate="DE" />

<word id="8" cont="的" pos="u" parent="9" relate="SBV" />

<word id="9" cont="是" pos="v" parent="0" relate="VOB" />

<word id="10" cont="统一" pos="n" parent="13" relate="ATT" />

<word id="11" cont="还是" pos="c" parent="12" relate="LAD" />

<word id="12" cont="分裂" pos="n" parent="10" relate="COO" />

<word id="13" cont="问题" pos="n" parent="9" relate="VOB" />

<word id="14" cont="。" pos="wp" parent="-2" relate="PUN" />

</sent>

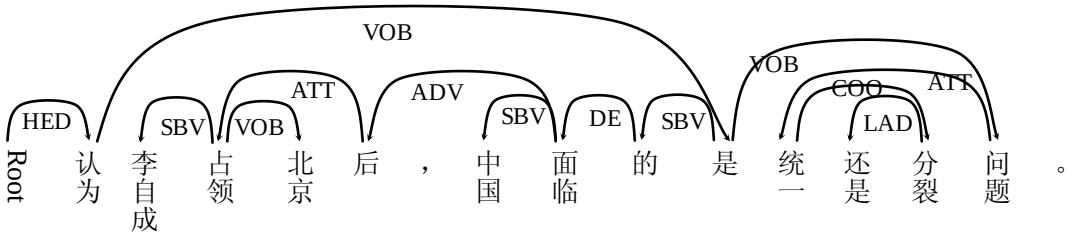


图 4.5 例句（5）标准依存分析图

其中“<sent”和“</sent>”表示句子相关内容的开始和结束标记，其中包含句

子在文本中的位置“id”、句子的具体内容“cont”以及句子中每个词语的详细信息——“word”表示词语，其后包含该词在句子中的位置“id”、词语内容“cont”、词性“pos”、依存句法分析结果的父节点“parent”以及依存关系“relate”。其对应的依存句法分析结果如图 4.5 所示：

步骤二：从步骤一生成的语料中抽取句子及其分词和词性标注信息作为新的语料，以克服因分词和词性标注错误对依存句法分析的影响，即：

<sent id="9437" cont="认为李自成占领北京后，中国面临的是统一还是分裂问题。">

<word id="0" cont="认为" pos="v" />

<word id="1" cont="李自成" pos="nh" />

<word id="2" cont="占领" pos="v" />

<word id="3" cont="北京" pos="ns" />

<word id="4" cont="后" pos="nd" />

<word id="5" cont="，" pos="wp" />

<word id="6" cont="中国" pos="ns" />

<word id="7" cont="面临" pos="v" />

<word id="8" cont="的" pos="u" />

<word id="9" cont="是" pos="v" />

<word id="10" cont="统一" pos="n" />

<word id="11" cont="还是" pos="c" />

<word id="12" cont="分裂" pos="n" />

<word id="13" cont="问题" pos="n" />

<word id="14" cont="。" pos="wp" />

</sent>

步骤三：以步骤二的结果作为输入语料，对其进行依存句法分析，得到 LTP 的句法分析结果，即

<sent id="0" cont="认为李自成占领北京后，中国面临的是统一还是分裂问题。">

<word id="0" cont="认为" pos="v" parent="-1" relate="HED" />

<word id="1" cont="李自成" pos="nh" parent="2" relate="SBV" />

<word id="2" cont="占领" pos="v" parent="4" relate="ATT" />

<word id="3" cont="北京" pos="ns" parent="2" relate="VOB" />

<word id="4" cont="后" pos="nd" parent="7" relate="ADV" />
<word id="5" cont="，" pos="wp" parent="-2" relate="WP" />
<word id="6" cont="中国" pos="ns" parent="7" relate="SBV" />
<word id="7" cont="面临" pos="v" parent="8" relate="DE" />
<word id="8" cont="的" pos="u" parent="10" relate="ATT" />
<word id="9" cont="是" pos="v" parent="10" relate="ATT" />
<word id="10" cont="统一" pos="n" parent="0" relate="VOB" />
<word id="11" cont="还是" pos="c" parent="13" relate="LAD" />
<word id="12" cont="分裂" pos="n" parent="13" relate="ATT" />
<word id="13" cont="问题" pos="n" parent="10" relate="COO" />
<word id="14" cont="。" pos="wp" parent="-2" relate="WP" />
</sent>

其对应的依存分析图如图 4.6 所示。

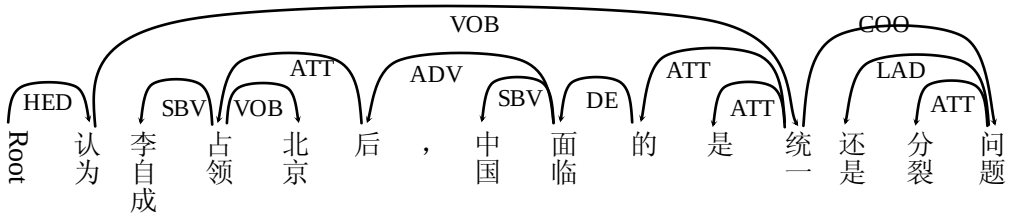


图 4. 6 例句（5）LTP 依存分析图

步骤四：为了之后对句子中的连词进行用法标注，对步骤二生成的语料进行格式转化，将其转化为北京大学《人民日报》中分词和词性标注形式，即：

认为/v 李自成/nh 占领/v 北京/ns 后/nd ，/wp 中国/ns 面临/v 的/u 是/v 统一/n 还是/c 分裂/n 问题/n 。/wp

步骤五：因为依存树库的词性标注采用的是 863 词性标注集，与北京大学的词性标注集有一些不同[49]，而上下文的词性信息对虚词用法标注有影响，因此，需要将步骤四产生的语料进行词性转换，即：

认为/v 李自成/**nr** 占领/v 北京/ns 后/**f** ，/**wd** 中国/ns 面临/v 的/**ud** 是/v 统一/n 还是/c 分裂/n 问题/n 。/**wj**

步骤六：对步骤五产生语料中的连词进行用法标注，其标注结果如下：

认为/v 李自成/nr 占领/v 北京/ns 后/f ,/wd 中国/ns 面临/v 的/ud 是/v 统一/n **还是/c<c_hai2shi4_1a>** 分裂/n 问题/n 。/wj

步骤七：对步骤六的语料进行连词结构短语识别[48]，标注句子中的连词结构短语，即：

认为/v 李自成/nr 占领/v 北京/ns 后/f ，/wd 中国/ns 面临/v 的/ud 是/v **<CP_bl> 统一/n 还是/c<c_hai2shi4_1a> 分裂/n </CP_bl>** 问题/n 。/wj

其中<CP_bl>和</CP_bl>分别表示并列结构短语的起始和结束位置，即并列结构短语为“统一还是分裂”。

步骤八：根据识别的连词结构短语对步骤三生成的依存句法分析的结果进行优化。因为本节只考查依存关系中的并列关系，所以只优化连词结构短语包含的词语之间的依存关系。对步骤三的结果进行优化的具体方法如下：

1) 找到连词结构短语中连词的前一个词语，该词即为并列关系中的一个并列成分，如“统一”；
2) 继续向前查找连词结构短语中的并列标记——顿号，如果没有顿号，转去执行 3)；如果有顿号，则顿号前的词语也属于并列成分，继续查找直至连词结构短语的第一个词语，最后查找到的那个词语便为第一个并列成分，也是其他并列成分的父节点；
3) 查找连词结构短语中的最后一个词语为最后一个并列成分，如“分裂”；
4) 将除第一个并列成分外的所有并列成分的“parent”重置为第一个并列成分的“id”，“relate”重置为“COO”，如“分裂_10_COO”；
5) 将连词的“parent”重置为最后一个并列成分的“id”，“relate”重置为“LAD”，如“还是_12_LAD”。

例句（5）的优化后的结果如图 4.7 所示：

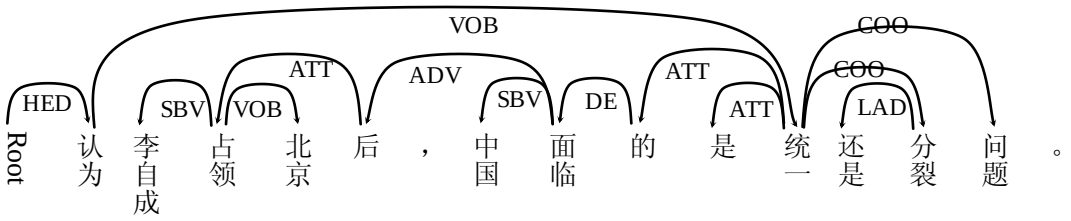


图 4.7 例句（5）优化后的依存分析图

词语“统一”和“分裂”是并列关系，且为了满足左核心原则，将“统一”作为“分裂”的父节点，并列标记“还是”依存于其后的并列成分“分裂”。

对比图 4.5、图 4.6、图 4.7，发现通过连词用法标注及连词结构短语的标注，

不仅正确识别了并列关系 COO，同时也可以影响前附加关系 LAD 的识别效果。

以连词“和”、“及”、“与”、“还是”为例采用上述步骤对抽取的语料进行实验，实验结果如表 4.3 所示，“LTP + usage”表示加入虚词用法信息后的实验结果。当加入连词用法后，这些句子中包含的并列关系 LAS 提高了 3.43%，UAS 提高 2.29%，说明连词的用法可以应用到依存句法分析中，提高并列关系的识别效果。

表 4. 3 加入连词用法前后 COO 的识别结果对比

词语	LTP		LTP + usage	
	LAS(%)	UAS(%)	LAS(%)	UAS(%)
和	96.21	97.15	99.05	99.05
及	85.00	90.00	90.00	90.00
与	96.30	96.30	100.00	100.00
还是	75.00	75.00	100.00	100.00
总计	95.04	96.18	98.47	98.47

对实验的语料进行分析后发现，因为依存句法分析中的每个词语只能有一个父节点，为了满足这个条件，在对其中的并列关系进行优化时，也可能对其他的依存关系弧进行修改。也就是说，加入虚词用法不仅会影响并列关系的识别情况，对句子中的其他依存关系的识别情况也存在影响，如例句（5）中的前附加关系 LAD。

4.3 其他虚词的用法在依存句法分析中的应用

在上节的研究分析中，主要考察了连词用法在依存句法分析的应用，虚词中的其他词性的用法也可以影响依存句法分析中依存关系的识别情况，如副词用法、介词用法等。

首先分析以下两个例句：

（6）**最**矮的那个人是小明。

（7）个子**最**矮的那个人是小明。

这两个例句的依存句法分析结果分别如图 4.8 和图 4.9a 所示，两个句子中的词语“最”都为副词，并且都修饰其后的形容词“矮”构成定中结构 ADV，然后与助词“的”构成“的”字结构 DE，作为主语“人”的修饰成分（与“人”构成定中关系 ATT）。

对两个例句进行分析后发现，例句（6）中副词“最”与形容词“矮”一起构成“最矮”修饰的是主语“人”，而例句（7）中副词“最”与形容词“矮”一起构成“最矮”作为补语修饰主语“个子”，然后与“的”构成新的修饰成分“个子最矮的”修饰主语“人”，这与图 4.9a 的结果不同，即例句（7）的 LTP 依存句法分析结果存在错误，标准的句法分析结果如图 4.9b 所示。

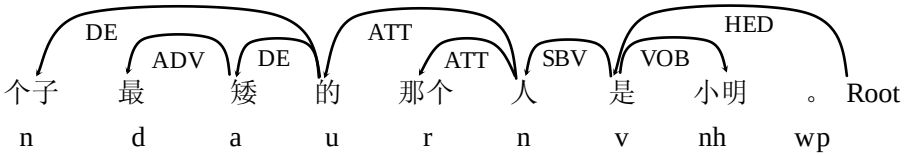
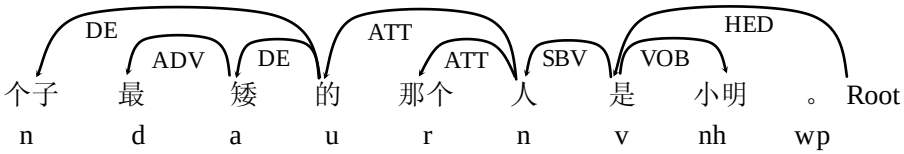
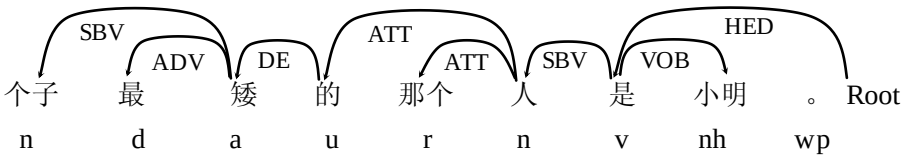


图 4.8 例句（6）的句法分析结果



（a）



（b）

图 4.9 例句（7）的句法分析结果

虽然例句（7）中的“最矮”暗含了“个子最矮”的意思，即两个例句表达的意思相同，但在进行依存句法分析时，主要分析的是两个词语之间的关系，如果将例句（7）按照图 4.9a 进行依存句法分析，就会导致名词“个子”与“的”组成“的”字结构 DE 作为定语修饰中心语“人”，短语“个子的人”显然不合中文的表达习惯。因此一定要将“矮”与其依存关系比较紧密的词语“个子”先构成主谓关系，然后在与其依存关系相对较远的词语“人”构成定中关系 ATT。

上述内容主要是依据中文的表达习惯和依存句法分析的原则来分析的，考虑到两个例句中都存在副词“最”，可以从副词的用法角度进行分析。

两个例句中副词“最”的用法标注情况如下：

（6）最/d<d_zui4_1aa> 矮/a 的/ud 那个/rz 人/n 是/v 小明/nr 。/wj

（7）个子/n 最/d<d_zui4_1ab> 矮/a 的/ud 那个/rz 人/n 是/v 小明/nr 。/wj

两个例句中的副词“最”用法标注不同，其中用法<d_zui4_1aa>后跟形容词，主要修饰名词，一般与“的”搭配，用法<d_zui4_1ab>后也跟形容词，可以做谓语或者补语。也就是说，例句（7）中副词“最”的用法标注“最/d<d_zui4_1ab>”可以将“个子”和“最矮”结合在一起，作为一个整体再参与其他词语的句法分析。这样可以同时影响主谓关系 SBV 和“的”字结构 DE 两种依存关系的识别效果，由此，虚词的用法信息可以作为依存句法分析的影响因素。

4.4 本章小结

本章主要分析了 LTP 中依存句法分析的 24 种依存关系的识别情况，并在虚词用法识别的基础上，讨论汉语虚词用法信息对依存句法分析的影响。包含并列关系的句子中多数含有虚词，并且并列关系是依存句法分析的难点，因此着重考虑了虚词用法在并列关系识别中的应用和影响。实验表明，考虑到虚词用法的因素后，并列关系的识别情况有明显提高，同时也对与并列关系相关的依存关系有一定的影响，肯定了虚词用法对并列关系的影响。在此基础上，简单探讨了其他虚词的用法信息对依存句法分析其他依存关系的影响。

5 结论与展望

5.1 结论

本文首先在现代汉语虚词相关研究的基础上对虚词中副词的用法进行了大量的分析和研究，并完善了“三位一体”的现代汉语虚词用法知识库，包括虚词用法语料库、虚词用法规则库和虚词用法词典。然后采用基于规则的方法对现代汉语中的副词用法进行了自动识别研究，不断优化用法的规则库以提高规则的准确率。这些规则是通过对语料的分析总结出来的规律以及语言学知识编写的，在一定程度上可以对各种用法进行区分，但并不能完全包含各个用法的信息，如语气、感情色彩等等，因此又采用了基于统计的方法对副词的用法进行了自动识别研究。本文分别采用了支持向量机、最大熵和条件随机场三种统计模型进行了实验，实验结果表明支持向量机模型的效果最好，且三种统计模型的识别效果在总体上优于基于规则的方法，但对于某些用法来说，基于统计方法识别效果比规则方法差。在进行统计学习时，用法出现的次数越多，识别的准确率就越高，但一些用法在语料中出现次数较少而无法进行统计学习，这降低了基于统计方法的识别效果。之后分析基于规则和基于统计两种方法的优点，考虑将两种方法结合起来，使用规则和统计相结合的方法对副词的用法进行自动识别研究。实验结果表明，规则和统计相结合的方法优于单独使用规则方法和单独使用统计方法的识别效果。

在虚词用法自动识别的基础上，本文将现代汉语虚词用法应用到依存句法分析研究中。通过对哈尔滨工业大学实现的依存句法分析器进行详细的分析和考察，总结 24 种依存关系的识别情况，发现其中的并列关系识别效果较差。经分析发现并列关系 COO 一般与连词一起出现，如“和”、“并”、“及”、“与”等，这些连词在句子中作为并列成分的连接词。基于以上的分析，本文根据这些连词的用法识别出句子中的连词结构短语，然后根据连词结构短语判断出短语中的各个并列成分，对依存句法分析的结果进行优化，进而改进依存句法分析中并列关系的识别情况。实验结果表明采用这种方法，依存句法分析中并列关系的识别准确率提高了。最后分析探讨了其它虚词的用法信息对依存句法分析的影响。

5.2 展望

虚词是现代汉语中的重要组成部分，对现代汉语虚词用法的研究也是汉语研究的重点，对虚词用法的应用研究才刚刚开始，还有很大的研究空间。在以后的研究中，可以考虑在以下几个方面进行探索：

在基于规则的用法自动识别研究中，规则的好坏直接关系到用法的识别效果，所以要不断的对规则进行优化，或者提出更好的优化方法，从而提高规则的质量。

在基于统计的用法自动识别研究中，各个统计模型特征模版的设计非常重要，针对那些出现频率较少的用法识别效果较差，可以为其设计特殊的特征模板，或者为这些用法设计特殊的标记，在进行统计学习前先进行特殊的处理，以提高其识别效果。

在规则和统计相结合的用法自动识别中，本文提出的相结合思想不一定适合所有的虚词。通过对各种虚词进行分析，寻找更能体现规则和统计方法优点的结合信息，或者根据用法个数的不同、各用法分布率情况等指标将词语进行分类，针对每种类型的词语设计出相应的结合方法。

现代汉语虚词用法知识库的应用研究刚刚起步，本文仅仅针对部分虚词用法对依存句法分析中并列关系的影响进行了探讨，下一步可以具体分析探讨虚词用法对句法分析中的其他依存关系的影响，以及虚词用法在自然语言处理其他领域中的应用。

参考文献

[1] 董振东，董强，郝长伶. 下一站在哪里[J].中文信息学报，2011，25（6）：3～11
[2] 陆剑明，马真. 现代汉语虚词散论（修订版）[M]. 北京：语文出版社，1999
[3] 俞士汶，朱学锋，刘云. 现代汉语广义虚词知识库的建设[J]. 汉语语言与计算学报，2003，2（1）：89～98
[4] 昝红英，朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建[J]. 当代语言学，2009，11（2）：124～135
[5] 刘锐. 基于规则的现代汉语副词用法自动识别研究[D]. [硕士学位论文]. 郑州：郑州大学，2009
[6] 张军珲. 基于统计的常用汉语副词用法自动识别研究[D]. [硕士学位论文]. 郑州：郑州大学，2010
[7] 袁应成. 基于用法属性的现代汉语介词短语边界识别研究[D]. [硕士学位论文]. 郑州：郑州大学，2011
[8] 韩英杰，昝红英，张坤丽等. 基于规则的现代汉语常用助词用法自动识别[J]. 计算机应用，2011，31（12）：3271～3274
[9] 周溢辉. 现代汉语语气词用法自动识别研究[D]. [硕士学位论文]. 郑州：郑州大学，2011
[10] 周丽娟. 现代汉语连词用法的自动识别及应用研究[D]. [硕士学位论文]. 郑州：郑州大学，2012
[11] 张坤丽，赵丹，昝红英等. 常用现代汉语副词用法自动识别研究[J].中文信息学报，2012，26（6）：65～71
[12] 袁应成，昝红英，张坤丽等. 基于规则的虚词用法自动标注算法设计与系统实现[C]. 第十一届汉语词汇语义学会议论文集（CLSW2010）. 苏州：苏州大学，2010，163～169
[13] 俞士汶，段慧明，朱学锋等. 北京大学现代汉语语料库基本加工规范[J].中文信息学报，2002，16（5）：58～65
[14] Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. Proceedings of the 18th ICML-01,2001: 282～289.
[15] E.T.Jaynes. Information Theory and Statistical Mechanics [J]. Physics Reviews, 1957: 620～630
[16] Vapnik V N. Statistical Learning Theory [M]. Wiley-Interscience Publication, 1998
[17] 昝红英，张坤丽，柴玉梅等. 现代汉语虚词知识库的研究[J]. 中文信息学报， 2007，21（5）：107～111
[18] 俞士汶，朱学锋，刘云. 面向自然语言理解的汉语虚词研究[C].见：嘎日迪，吾守尔编. 第十一届全国民族语言文字信息学术研讨会论文集. 北京：西苑出版社，2007，270～279
[19] 陈火旺. 程序设计语言编译原理（第三版）[M]. 北京：国防工业出版社，2000

[20] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008
[21] 刘海涛. 依存句法的理论与实践[M]. 北京：科学出版社，2009
[22] Zhou Q. A statistics-based Chinese Parser[C]. In Proceedings of the 5th Workshop on Very Large Corpora, New York, 1997:4～15
[23] 李正华. 依存句法分析统计模型及树库转化研究[D]. [硕士学位论文]. 哈尔滨：哈尔滨工业大学，2008
[24] Tesniere L. Elements de syntaxe structurale[M]. Paris: Klincksieck. 1959.
[25] Robinson J J. Dependency structures and transformational rules [J]. Language. 1970. 46(2): 259～285
[26] 冯志伟. 形式语言理论[J].计算机科学，1979（1）：34～57
[27] Eisner. Three new probabilistic models for dependency parsing: an exploration[C]. Coling.1996.340～345
[28] Collins. A new statistical parser based on bigram lexical dependencies[C]. ACL.1996.184～191
[29] Ryan McDonald, Koby Crammer, Fernando Pereira. Online Large-Margin Training of Dependency Parsers[C]. Association for Computational Linguistics (ACL). 2005. 91～98
[30] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[C]. In Proceedings of the Coling 2010.13～16
[31] Ryan McDonald, Fernando Pereira. Non-projective Dependency Parsing using Spanning Tree Algorithms[C]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005.523～530
[32] Ryan McDonald, Fernando Pereira. Online Learning of Approximate Dependency Paring Algorithms[C]. European Association for Computational Linguistics (EACL). 2006.81～88
[33] Zhou, M. A block2based robust dependency parser for unrestricted Chinese text[C]. Proceedings of 2nd Chinese Language Processing Workshop. ACL.2000.224～230
[34] 王蕾. 基于统计方法的汉语长句依存句法分析[D]. [硕士学位论文].青岛：中国海洋大学，2009
[35] 袁里驰. 基于依存关系的句法分析统计模型[J]. 中南大学学报（自然科学版），2009，40（6）：1630～1635
[36] 袁里驰. 基于词聚类的依存句法分析[J]. 中南大学学报（自然科学版），2011，42（7）：2023～2027
[37] 刘海涛，赵怿怡. 基于树库的汉语依存句法分析[J]. 模式识别与人工智能，2009，22（1）：17～21
[38] 左维松. 规则和统计相结合的篇章情感倾向性分析研究[D]. [硕士学位论文]. 郑州：郑州大学，2010
[39] 潘正高. 基于规则和统计相结合的中文命名实体识别研究[J]. 情报科学，2012，30（5）：709～712
[40] 刘锐，昝红英，张坤丽. 现代汉语副词用法的自动识别研究[J]. 计算机科学， 2008， 8（A）：172～174

[41] 昝红英，张军珲，朱学锋. 副词“就”的用法及其自动识别研究[J]. 中文信息学报，2010，24（5）:10～16

[42] 黄德根,焦世斗,周惠巍. 基于子词的双层 CRFs 中文分词[J]. 计算机研究与发展, 2010, 47（5）：962～968

[43] 应玉龙,李淼,乌达巴拉等. 基于条件随机场的蒙古语词性标注方法[J]. 计算机应用, 2010, 30（8）：2038～2040

[44] 史海峰. 基于 CRF 的中文命名实体识别研究[D]. [硕士学位论文].苏州: 苏州大学, 2010

[45] 李正华，车万翔，刘挺. 基于柱搜索的高阶依存句法分析[J]. 中文信息学报，2010，24（1）:37～41

[46] Wanxiang Che, Min Zhang, Haizhou Li. Utilizing Dependency Languege Models for Graph-based Dependency[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012：213～222

[47] Ting Liu, Jinshan Ma, Sheng Li. Building a Dependency Treebank for Improving Chinese Parser[J]. Journal of Chinese Language and Computing, 2006,16（4）:207～224

[48] 昝红英，周丽娟，张坤丽. 基于用法的现代汉语连词结构短语识别研究[J].中文信息学报，2012，23（6）：72～78

[49] 马金山，基于统计方法的汉语依存句法分析研究[D]. [博士学位论文]. 哈尔滨：哈尔滨工业大学，2007

个人简历、在学期间发表的学术论文及研究成果

个人简历

张静杰，女，1986 年 10 月 08 日出生，河南省获嘉县人。
2006 年 09 月考入郑州大学计算机科学与技术专业，2010 年 06 月本科毕业并获得工学学士学位；
2010 年 09 月考入郑州大学计算机软件与理论专业，工学硕士研究生在读。

在学期间发表的学术论文

张静杰，昝红英. 副词“都”用法自动识别研究[J].北京大学学报（自然科学版），2013，49（1）

致谢

自从 2010 年 6 月份开始，我便有幸参与了郑州大学自然语言处理实验室虚词项目的研究工作。我感到非常荣幸，也非常珍惜大学毕业以后还可以有在大学校园里继续学习深造的机会。三年的时间转瞬即逝，过去的美好时光令人难忘。在郑州大学的学习和生活中，我得到了很多老师和同学的无私帮助，在此，我向所有关心和帮助过我的老师和同学们表示衷心的感谢和由衷的敬意。

首先非常感谢我的导师咎红英教授，三年多来在学习上和生活上都给了我无私的关怀和帮助。咎红英老师严谨的治学态度和平易近人的性格深深影响着我，正是由于她的认真负责才使我可以不断的进步。咎红英老师时时刻刻关注着我的学习情况，尤其在论文的开题、研究工作的开展以及论文的写作、修改、定稿等等整个过程都得到了咎老师的指导和帮助。她的博学求实、为人处事使我学习到了很多，这些不仅对我的研究生生活有很大的帮助，对我以后的工作生活也有很大的影响。能够在硕士学习期间遇到这样一位好老师，我感到非常的荣幸。

感谢郑州大学自然语言处理实验室的柴玉梅老师、穆玲玲老师、张坤丽老师、韩英杰老师、贾玉祥老师、赵丹老师在项目研究和生活上对我的帮助和指导，你们的每次指导都让我受益匪浅，非常感谢你们。

感谢郑州大学信息工程学院的周清雷老师、范明老师、叶阳东老师、庄雷老师、王黎明老师、赵东明老师、徐江峰老师等各位老师对我学习上的辅导，也感谢庞军老师和高明磊老师在硕士研究生阶段给我的各方面的帮助。

感谢朝夕相处的郑州大学自然语言处理实验室的成员：庞熠雅、张腾飞、梁猛杰、任丽君、李钰、范庆虎、娄鑫坡、贺娟、杨银涛、孙佳、吴泳刚、刘铭、李元一、李静毅、王浩海、刘一韬等在学习和生活上对我的帮助。在实验室这个温暖的大家庭里，我学到了很多，和大家一起度过的这段时间将是我以后生活中的美好回忆。

感谢我的家人对我的支持和鼓励。

最后，感谢参加论文评审和答辩的各位专家和老师，向你们致以深深的敬意。